



Reid, Stephen James (2012) *Trends of organic carbon in Scottish rivers and lochs*. MSc(R) thesis

<http://theses.gla.ac.uk/3379/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



## **Trends of Organic Carbon in Scottish Rivers and Lochs**

Stephen James Reid

*A Dissertation Submitted to the*

*University of Glasgow*

*for the degree of Master of Science*

School of Mathematics and Statistics

October 2011

© Stephen James Reid

# Abstract

In Scotland, the Scottish Environment Protection Agency (SEPA) is the regulatory agency responsible for monitoring water quality and reporting back to the Scottish and UK governments and the European community. In order to improve water quality in surface waters such as rivers and lochs, the European Parliament has established directives over the past twenty years outlining targets for nutrient levels and water quality status. Moxley (2010), states that the concentration of organic carbon in many Scottish rivers, has approximately doubled over the last twenty years, with soils being the most likely source. According to Moxley (2010), the rate of total organic carbon (TOC) increase, averaged across all sites with increasing concentrations, was 0.12 milligrams per litre per year (mg/l/y). This is an increase in TOC concentration of nearly 2.5 mg/l over a twenty year period. Consequently, the behaviour of organic carbon in Scottish rivers and lochs has become of interest and is the focus of analysis within this thesis.

Chapter 1 introduces organic carbon, providing an insight into observed trends in the United Kingdom, but also, other parts of the world. Furthermore, Chapter 1 discusses environmental and physical factors which are thought to be associated with changing levels in organic carbon. Moreover, Chapter 1 provides a description of the data and sampling techniques which have been used. The exploratory analysis in Chapter 2 reveals that the log TOC levels in rivers and lochs have been increasing up until the early 2000's, and that the log TOC

follows a seasonal pattern. Furthermore, the exploratory analysis reveals the high level of association between total organic carbon and dissolved organic carbon. The exploratory analysis also highlighted issues with the covariates; therefore Chapter 2 explores suitable methods of dealing with values at the limit of detection, as well as appropriately imputing missing values.

Chapter 3 explores log TOC at a selection of river and loch sites, and the relationship between log TOC and covariates at each site in detail. In addition, Chapter 3 explores the use of different regression techniques (e.g. linear and additive modelling) in order to appropriately capture the behaviour of log TOC at each site. Chapter 4 progresses from investigating and modelling individual sites, to exploring sites which are connected in some manner. Chapter 4 considers the behaviour of log TOC in sites which are part of the River Dee network, where the distance between each site and how the river flows between each of the sites had to be taken into consideration. Chapter 4 investigates the behaviour of log TOC across the river network over time and space visually; but, also explores appropriate modelling techniques which suitably capture the behaviour of log TOC over time and space, taking into consideration suitable covariates to plausibly explain the observed trends.

Chapter 5 addresses the main theme of the thesis: coherency is defined and explored there. A literature review was conducted to consider possible methods of measuring coherency. The seasonal Mann-Kendall test was found to be an appropriate method of measuring the heterogeneity of a group of sites; and dynamic factor analysis was found to be an effective technique of identifying common trends in a group of sites; hence, these methods were applied in Chapter 5 to measure the level of coherency between sites in the River Dee network, but also, sites located in the same Scottish region. Progressing from the analysis carried out in Chapter 5, Chapter 6 aims to appropriately model the levels of log TOC in Scottish regions, taking into account time and space, but also, possible covariates thought to be driving such trends. Finally, Chapter 7 provides a summary of the findings, and discusses limitations of the study and possible areas of future research.



# Acknowledgements

I would like to take this opportunity to thank my supervisor Prof. Marian Scott for her priceless guidance and support throughout the year, which has contributed to a thoroughly enjoyable MSc experience. Furthermore, I would like to thank the Scottish Environment Protection Agency for providing the data for this project; in particular, I would like to thank Mark Hallard and Janet Moxley for taking the time to assist myself throughout the year. In addition, I gratefully acknowledge the funding from ISD which has allowed me to undertake this project.

I would like to thank the post graduates (Ruth Haggarty, David O'Donnell, Alastair Rushworth) and Dr Claire Ferguson for their friendly nature and willingness to help throughout the year. Thanks are also due to my fellow MSc colleagues (Kathryn McNeil, Laura Allison, Mhairi Kerr, Collette Letham, Greg Halbert, Andisheh Bakhshi) and Martin McKean for making this year so memorable.

Finally, to my mum, dad, brother and sister, thank you! Your love, encouragement support and sense of humour, has played a key role in my life achievements and successful university career.

# Contents

Chapter 1 .....	1
Introduction.....	1
1.1 Background and Motivation for Research .....	1
1.2 Factors Driving Trends.....	4
1.3 SEPA: Sampling and Measuring of Data .....	7
1.3.1 SEPA: Measuring DOC and TOC Concentrations .....	7
1.3.2 Covariates .....	9
1.4 Missing Data and Time Series .....	10
1.5 Overview of Thesis .....	12
Chapter 2 .....	14
Exploring Trends, Seasonality and Relationships .....	14
2.1 Initial Impression of Total Organic Carbon .....	15
2.2 Specific Trends of Log TOC in Rivers and Lochs .....	18
2.3 Seasonality of Total Organic Carbon in Rivers and Lochs .....	23
2.4 Relationship between TOC and DOC.....	26
2.5 Further Exploratory Analysis – Log TOC Relationships With Covariates .....	29
2.6 Values at the limit of detection: Regression on Order Statistics (ROS) .....	34
2.7 Predicting Temperature .....	36
2.8 Conclusions.....	37
Chapter 3 .....	39

Modelling Trend, Seasonality and Covariates at Sites .....	39
3.1 Initial Impression of the sites .....	40
3.2 Relationship Between Log TOC and Covariates .....	44
3.3 Modelling Log TOC At Each Site .....	47
3.3.1 Harmonic Regression .....	47
3.3.2 Auto-Correlation of Residuals .....	51
3.3.3 Fitting Multiple Linear Regression Models .....	52
3.3.4 Additive Models and Non-Parametric Regression .....	56
3.3.5 Fitting Additive Models to Sites .....	58
3.4 Choosing The ‘Best’ Model: Linear or Additive? .....	66
3.5 Conclusion .....	68
Chapter 4 .....	71
River Networks .....	71
4.1 Initial Impression of the Sites Along the Main .....	72
Channel (i.e. The River Dee) .....	72
4.2 Modelling Each Site Along the Main Channel (i.e. The River Dee) .....	77
4.3 Modelling the Levels of Log TOC on the Main Channel: Finding a Global Model .....	86
4.3.1 Global Modelling: Generalized Additive Mixed Models (GAMM’s) .....	87
4.4 The River Dee Network .....	95
4.4.1 Trends, Seasonality and Relationships .....	97
4.4.2 Measures of Spatial Dependence .....	100
4.4.3 Flow Connected Sites .....	104
4.4.4 Moving Average Constructions and Valid Covariances .....	105
4.4.5 Modelling the River Network .....	107

4.4.6 Visualising Trend Over Space .....	113
4.4.7 Modelling the River Dee Network: Non-Parametric Regression Over Time and Space.....	116
4.4.8 Conclusions of the River Dee Network.....	123
Chapter 5.....	127
Coherency .....	127
5.1 Literature Review .....	128
5.2 Methodology .....	137
5.2.1 Seasonal Mann Kendall Test.....	137
5.2.2 Dynamic Factor Analysis .....	140
5.3 Applications of Methodology: River Dee Network .....	143
5.3.1 Applying the Seasonal Mann-Kendall Test to the River Dee Network.....	143
5.3.2 Applying Dynamic Factor Analysis to the River Dee Network.....	146
5.4 River Dee Network Conclusion.....	149
5.5 Scottish Regions .....	150
5.5.1 Initial Impression of Regions .....	151
5.5.2 Applying the Seasonal Mann-Kendall Test to Scottish Regions .....	157
5.5.3 Applying the Dynamic Factor Analysis to the Scottish Regions .....	159
5.6 Conclusion .....	166
Chapter 6.....	169
Modelling Log TOC, Over Time and Space in Scottish Regions.....	169
6.1 Modelling Trend and Seasonality.....	170
6.2 Modelling Trend, Seasonality and Covariates.....	174
6.3 Conclusion .....	187

Chapter 7 .....	189
Discussions and Conclusions.....	189
7.1 Summary .....	189
7.2 Limitations of the Study and Future Work.....	197
7.3 Conclusion .....	199
Bibliography .....	201

# List of Figures

Figure 1.1.1: Flow chart of Total Carbon (Sepa Chemistry <sup>1</sup> , <i>ES-INR-P-004</i> ) .....	3
Figure 1.4.1: Summary of length of time series at river (a) and loch (b) sites with regards to the TOC data available.....	11
Figure 2.1.1: Scatter plots of TOC against year at river (a) and lochs sites (d); Log TOC against year at river (b) and loch (e) sites; Log TOC against year with outliers removed at river(c) and loch (f) sites. ....	16
Figure 2.2.1: Scatter Plots of Log TOC against Year at a selection of river sites, with 10 Sites displayed on each plot.....	19
Figure 2.2.2: Scatter Plots of Log TOC against year at a selection of individual river sites with a lowess curve fitted. ....	20
Figure 2.2.3: Scatter Plots of Log TOC against year for a selection of loch sites. 10 sites displayed on each plot.....	21
Figure 2.2.4: Scatter Plots of Log TOC against year for a selection of individual loch sites with a lowess curve fitted. ....	22
Figure 2.3.1: Scatter Plots of Log TOC against day of the year for a selection of river sites (a)-(c); and plots of individual river sites with a loess curve fitted (d)-(e).....	24
Figure 2.3.2: Scatter Plots of Log TOC against day of the year for a selection of loch sites (a)-(c); and plots of individual river loch with a loess curve fitted (d)-(e). ....	25
Figure 2.4 1: Scatter plots of log TOC against log DOC at a selection of river (a) and loch (b) sites. ....	27
Figure 2.5.1: Time series of river flow at 49 sites with (a) and without (b) the use of the log transformation. ....	29

Figure 2.5.2: Scatter plots of log TOC against temperature (a), log flow (b), pH (c), log alkalinity (d), log sulphate (e) and log nitrate (f) at a selection of river sites.....	32
Figure 2.5.3: Scatter plots of log TOC against temperature (a), pH (b), log alkalinity (c), log sulphate (d) and log nitrate (e) at a selection of loch sites.....	33
Figure 2.6.1: Time series of log nitrate (mg/l) with and without the ROS computation at the River Muick – Allt Darrarie.....	35
Figure 2.7.1: Time series of log TOC (mg/l) against temperature (degrees Celsius) with and without predicted missing temperature values.....	36
Figure 3.1.1: Time series of log TOC (mg/l) at the three river sites [(a),(c) and (e)]; and seasonality plots of the three river sites [(b),(d) and (e)] with regards to log TOC levels. ....	42
Figure 3.1.2: Time series of log TOC (mg/l) at the three loch sites [(a),(c) and (e)]; and seasonality plots of the three loch sites [(b),(d) and (e)] with regards to log TOC levels. ....	43
Figure 3.2.1: Log TOC plotted against temperature (a), log alkalinity (b), pH (c), log nitrate (d), log sulphate (e) and log flow (f) [Callater Burn only] at the river sites.....	45
Figure 3.2.2: Log TOC plotted against temperature (a), log alkalinity (b), pH (c), log nitrate (d), log sulphate (e) at each of the loch sites.....	46
Figure 3.3.1.1: Trend and Seasonality Models fitted to the time series plots at the sites Callater Burn (a) and Loch Kilbirnie - Beith (b). ....	49
Figure 3.3.21: Auto-Correlation Function (a) and Partial Auto-Correlation Function (b) plots of the residuals from the trend and seasonality model fitted to Callater Burn.....	52
Figure 3.3.3.1: Residuals vs Fitted values plots for the final linear models fitted to Callater Burn(a), Loch Naver (b), Dall Bridge (c), Loch Kilbirnie (d), Tweed above Gala Waterfoot (e) and Loch Lomond (f).....	55

Figure 3.3.5.1: Effect plots of additive model fitted at: the River Tweed above Gala Waterfoot [(a)- (c)]; and Loch Kilbirnie [(d) and (e)]. .....	61
Figure 3.3.5.2: Residuals vs Fitted values plots for the final additive models fitted to Callater Burn(a), Loch Naver (b), Dall Bridge (c), Loch Kilbirnie (d), Tweed above Gala Waterfoot (e) and Loch Lomond (f).....	65
Figure 4.1: Location of the River Dee .....	72
Figure 4.1.1: Log TOC plotted against Year (a), Month (b), Log Alkalinity (c), Log Flow (d) and Log Sulphate (e) at the 5 river sites situated on the River Dee.....	75
Figure 4.1.2: Log TOC plotted against Temperature (a), pH (b) and log nitrate (c) at the 5 river sites situated on the River Dee. ....	76
Figure 4.2.1: Time series of Log TOC at sites 1 (a) and 2 (b) on the River Dee, with the corresponding trend and seasonality model fitted to each plot.....	78
Figure 4.2.2: Residuals vs Fitted Values plotted for the final linear models fitted to the sites Bridge of Dee (a), Milltimber (b), Potarch Bridge (c) and Linn of Dee (d). ....	80
Figure 4.2.3: A selection of effect plots from the final additive models fitted to sites Bridge of Dee (a), Milltimber (b), Potarch Bridge (c) and Linn of Dee (d). ....	82
Figure 4.3.1: Time series of log TOC levels at the 4 River Dee sites .....	87
Figure 4.3.1.1: Cross-Correlation Function plots of sites: 1 and 2 (a), 2 and 3 (b), using the residuals from the trend and seasonal linear models fitted in Section 4.2; Plot of lag 0 auto-correlation coefficients from CCF's (c). ....	90
Figure 4.3.1.2: Year (a), Month (b), Log Alkalinity (c), Log Flow (d) and Log Sulphate (e) effect plots of the GAMM model fitted to the River Dee.....	93
Figure 4.4.1 (right): Portion of the River Dee network plotted in blue and locations of 13 River Dee network sites marked in red on Figure 4.4.1 (a); and the corresponding 'site number' of each site is stated on Figure 4.4.1 (b).....	96



Figure 4.4.1.1: The trend (a) and seasonality (b) of log TOC at the thirteen sites; log TOC against pH (c) and log Alkalinity (d) at the thirteen sites. ....	98
Figure 4.4.1.2: Log TOC against temperature (a), log sulphate (b), log nitrate and log flow (d) at the thirteen sites. ....	99
Figure 4.4.2.1: Variograms of 13 River Dee network sites, using Euclidean.....	103
Figure 4.4.3.1: Directed Acyclic Graph used to express sites which are flow- connected across a network.....	104
Figure 4.4.5.1: Smooth estimates of 13 known locations and 217 new locations across the RiverDee Network, using Euclidean distance with the smoothing parameter $h = 5$ (a), 10 (b), 15 (c) and 20 (d). ....	111
Figure 4.4.5.2: Estimates of 13 known locations and 217 new locations across the RiverDee Network, using River distance with the smoothing parameter $h = 5$ (a), 10 (b), 15 (c) and 20 (d). ....	112
Figure 4.4.6.1: Estimates of 13 known locations and 217 new locations across the River Dee Network, using river distance (km) and month of March for the years 1990 (a), 1997 (b), 2000 (c) and 2009 (d). Note: site 5 is not included in the years 1997 and 2000. ....	115
Figure 4.4.7.1: Effect plots of the trend and seasonality GAM model fitted to the thirteen sites: Year (a), Month (b). 3D Trend and Seasonality plots of Callater Burn (c) and River Lui (d). Fitted values extracted from GAM model, for each site separately (e). ....	117
Figure 4.4.7.2: Smoothed mean partial residuals of the term $m_3(Location_i)$ for each of the thirteen sites (a); but, also the smoothed partial residuals of the other 217 new locations....	119
Figure 4.4.7.3: Effect plots of the final GAM model fitted to the thirteen sites: Log Alkalinity (a), pH (b) and log nitrate (c). Residuals vs Fitted values from the final GAM model fitted to the thirteen sites (c).....	122
Figure 5.3.1.1: The summer trend of the 13 River Dee network sites log TOC values.....	145

Figure 5.3.2.1: Factor loadings corresponding to the two common trends (a); Fitted values obtained by the DFA model with two common trends. ....	148
Figure 5.3.2.2: Residuals vs Fitted Values of the 13 time series from the final DFA model fitted. ....	148
Figure 5.5.1: Regions under investigation in Scotland. ....	151
Table 5.5.1.1: Summary of the river and lochs sites in each regio .....	152
Figure 5.5.1.1: Time series plots of log TOC in river sites at the Scottish regions: Argyll (a), Ayrshire (b), Borders (c) and Dumfries and Galloway (d). ....	153
Figure 5.5.1.2: Time series plot of log TOC in river sites in the Scottish regions: West Highlands (a), Perthshire (b) and Sutherland (c). ....	154
Figure 5.5.1.3: Time series plot of log TOC in loch sites at the Scottish regions: Dunbartonshire (a), West Highlands (b), Perthshire (c), Sutherland (d), Lewis (e) and Stirlingshire (f). ....	155
Figure 5.5.1.4: Seasonality of log TOC levels in the Argyll rivers (a); and seasonality of log TOC levels in the Argyll rivers during the year 2007 (b). ....	156
Figure 5.5.3.1: Selection of plots with the fitted values obtained from the final DFA models with one common trend. River sites located in the regions West Highlands (a), Sutherland (b) and Ayrshire (c). Loch sites located in the regions Stirling (d), Perthshire (e) and West Highlands (f). ....	161
Figure 5.5.3.2: A selection of residuals vs fitted values plots from the final DFA models fitted to the river sites in the regions: Ayrshire (a), Borders (b) and Sutherland (c). ....	164
Figure 5.5.3.3: A selection of residuals vs fitted values plots from the final DFA models fitted to the loch sites in the regions: Stirlingshire (a), Sutherland (b) and Dunbartonshire (c). ....	165

Figure 6.1.1: A selection of effects plots from the fitted trend and seasonality GAM models: year at the Argyll rivers (a), year at the Sutherland rivers (b), month at the Ayrshire rivers (c), year at the Dunbartonshire Lochs (d). Trend and seasonality 3D plots in the West Highland rivers site 1 (e) and site 2 (f). .....	173
Figure 6.2.1: A selection of effect plots from the final GAM models fitted- log nitrate in the Borders (a); log nitrate in Ayrshire (b); log alkalinity in Dumfries and Galloway (c); temperature in the West Highlands (d). .....	177
Figure 6.2.2: A selection of effect plots from the final GAM models fitted- pH in Dumfries and Galloway (a); log sulphate in the Borders (b); and annual rainfall in Ayrshire (c). .....	179
Figure 6.2.3: A selection of effect plots from the final GAM models fitted- pH in Stirlingshire (a); log sulphate in Stirlingshire (b); log alkalinity in Lewis (c); log nitrate in Perthshire (d). .....	182
Figure 6.2.4: A selection of effect plots from the final GAM models fitted- log nitrate in Lewis (a); temperature in Dunbartonshire (b); and annual rainfall in Lewis (c). .....	184
Figure 6.2.5: A selection of residuals vs fitted values plots extracted from the final GAM models fitted in the regions Ayrshire (a), Dumfries and Galloway (b), and Perthshire (c) [with regards to rivers]; and the regions Dunbartonshire (d), Perthshire (e) and Sutherland (f) [with regards to lochs]. .....	186

# List of Tables

Table 2.1.1: Summary of log TOC levels (mg/l) at river and loch sites.....	17
Table 2.4.1: Spearman's Rho coefficients for the correlation between a selection of river and loch sites.....	28
Table 3.3.1.1: Summary of the final trend and seasonality models fitted to the three river and three loch sites.....	49
Table 3.3.3.1: Summary of the final linear models fitted to each sites; and the significance of each term when included in the final linear models. ....	54
Table 3.3.5.1: The significance of each term, when included in the final additive semi-parametric model, at the River site Callater Burn.....	62
Table 3.3.5.2: The significance of each term when included in the final additive model at the River site Dall Bridge. ....	62
Table 3.3.5.3: The significance of each term when included in the final additive semi-parametric model at the River site Tweed above Gala Waterfoot. ....	63
Table 3.3.5.4: The significance of each term when included in the final additive model fitted at the site. ....	63
Table 3.3.5.5: The significance of each term when included in the final additive mode fitted at the site Loch Kilbirnie (Beith). ....	64

Table 3.3.5.6: The significance of each term when included in the final linear model at the site Loch Lomond (Creinch).....	64
Table 3.3.6.1: Comparison of final linear and additive models fitted to the river and loch sites using an Approximate F-test.....	67
Table 4.1.1: Summary of the 5 river sites under investigation situated on the River Dee .....	73
Table 4.2.1: The significance of each term when included in the final linear models fitted to each of the sites.....	79
Table 4.2.2: The significance of each term when included in the final additive semi-parametric model at the River site Bridge of Dee.....	83
Table 4.2.3: The significance of each term when included in the final additive semi-parametric model at the River site Milltimber.....	83
Table 4.2.4: The significance of each term when included in the final additive semi-parametric model at the River site Potarch Bridge. ....	84
Table 4.2.5: The significance of each term when included in the final additive semi-parametric model at the River site Linn of Dee.....	84
Table 4.2.6: Comparison of final linear and additive models fitted to the River Dee sites using an Approximate F-test.....	85
Table 4.3.1.1: Summary of the Final GAMM model fitted to the River Dee.....	94
Table 4.3.1.2: Summary of the Correlations between covariates and standard deviations in the GAMM model fitted to the River Dee.....	94
Table 4.4.1(left): List of the sites under investigation in the River Dee Network.....	96
Table 4.4.7.1: The significance of each term, when included in the trend and seasonality additive model, at the River Dee network. ....	118
Table 4.4.7.2: The significance of each term, when included in the final additive model, at the River Dee network. ....	121

Table 5.2.1.1: Summary of Seasonal Mann Kendall Test chi-square statistics and corresponding degrees of freedom.....	140
Table 5.3.2.1: Summary of Dynamic Factor Analysis models fitted to the 13 time series in the River Dee network. ....	147
Table 5.5.1.1: Summary of the river and lochs sites in each region. ....	152
Table 5.5.2.1: Summary of the Seasonal Mann Kendall tests performed on the specified regions.....	158
Table 5.5.3.1: Summary of the final DFA models fitted to each of the Scottish Regions, for rivers and lochs. ....	160
Table 5.5.3.2: Summary of the final DFA models fitted to each of the Scottish Regions, for rivers and lochs. ....	163
Table 6.1.1: Summary of the trend and seasonality GAM models fitted to the river sites located in the different regions of Scotland. ....	172
Table 6.1.2: Summary of the trend and seasonality GAM models fitted to the loch sites located in the different regions of Scotland. ....	172
Table 6.2.1: Summary of the final GAM models fitted to rivers in the specified regions.. ..	176
Table 6.2.2: Summary of the final GAM models fitted to lochs in the specified regions... ..	181

# Chapter 1

## Introduction

### 1.1 Background and Motivation for Research

In the past few decades, humanity's concern regarding the well-being of the environment has risen. The media continues to raise public awareness of each generation's "carbon footprint", while researchers and scientists investigate why the environment is changing. A key focus of study and research in Scotland, is the trend of organic carbon concentrations in rivers and lochs, particularly, Dissolved Organic Carbon (DOC) and Total Organic Carbon (TOC).

In Scotland, the Scottish Environment Protection Agency (SEPA) is the regulatory agency responsible for monitoring water quality and reporting back to the Scottish and UK Governments and the European community. In order to improve water quality in surface waters such as rivers and lochs, the European Parliament has established directives over the past twenty years outlining targets for nutrient levels and water quality status. For example,

legislation such as the Urban Waste Water Treatment Directive (UWWT, 1991) was introduced, which requires waste water for all cities or towns of more than 2,000 population equivalents discharging to freshwaters to be collected and treated appropriately according to the Directive. Furthermore, the European Union Water Framework Directive (WFD, European Parliament, 2000) has set aims to prevent further deterioration of Europe's water bodies by protecting and improving water quality in all aquatic ecosystems. In order to achieve a better quality of water, the Directive aims for progressive reduction in discharges and emissions of hazardous substances, as well as reducing the volume of man-made synthetic substances which are being introduced in some waters. In particular, the EU Water Framework outlines targets for nutrients levels in all rivers and requires control of discharges which cause limits to be exceeded. Over the years, European Water policy has undergone many amendments in an attempt to tighten the gaps in legislation and to ensure that our waters are as clean and safe as possible. The Water Framework Directive (WFD) monitoring is risk-based and focuses where there is likely to be a problem – the WFD offers suggestions about improving monitoring to maintain high standards of water quality ([http://ec.europa.eu/environment/water/water-framework/index\\_en.html](http://ec.europa.eu/environment/water/water-framework/index_en.html)).

Monitoring the organic carbon levels in rivers and lochs is important, as the carbon cycle is essential to the way in which ecosystems function and survive. Rivers play an important role in transporting carbon from the land to the oceans, with constant feedbacks to and from the atmospheric carbon pool (where the feedbacks involve sequestration or release of the greenhouse gases CO<sub>2</sub> and methane).

The carbon exported from catchments by rivers can be either organic or inorganic and in the phases of solid, solute or gas. All of this carbon has ultimately been drawn down from atmospheric CO<sub>2</sub> via the essential process of photosynthesis (Dixon and Turner, 1991). DOC plays an important role in aquatic systems by influencing light regime and nutrient supply, acidity, trace metal transport and potability. (Eimers et al., 2008)



Furthermore, monitoring carbon levels is important, as carbon lost from soils to water may be oxidised to produce carbon dioxide. Since Scotland has large stocks of organic carbon held in peaty and organic soils, estimated at 2735 metric megatons (MtC) in total (ECOSSE Report 2007), if it was all to be converted to carbon dioxide, it would be equal to 174 years of human emissions at current rates. Also, loss of organic carbon from soils also affects the quality of the soil – it leads to poor water holding capacity and impacts the soil’s ability to retain pollutants and nutrients. Loss of soil organic matter could increase run-off which in turn increases flood risk and pollutant content in water. The most obvious effect of an increase in the level of TOC is the darkening of water colour, which reduces available light and energy, particularly in deeper lochs. (Moxley, 2011)

SEPA published research generally refers to Dissolved Organic Carbon rather than Total Organic Carbon, similar to many other papers which have recently been published by other researchers – while, the focus of this thesis shall be on TOC. DOC is made up of a fraction of the water based carbon which can pass through a filter; however, TOC consists of particulate (NPOC) and purgeable (POC) carbon as Figure 1.1.1 displays. In the past, SEPA has generally monitored TOC in rivers and DOC in lochs. When rivers contain little suspended solid material DOC and TOC values are likely to be similar, but when sediment loadings are high, for example, in high flows, TOC values will be higher than DOC as TOC measurements include organic material bound to sediments (Moxley, 2011).

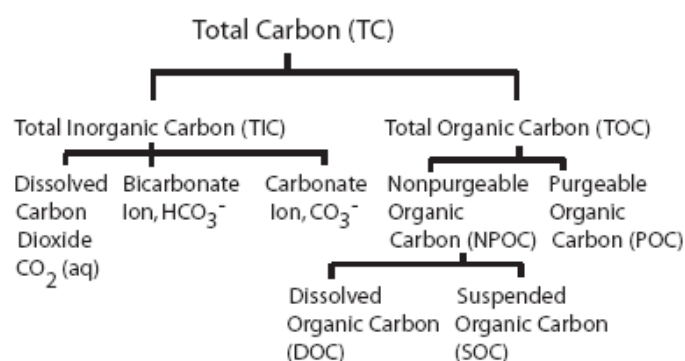


Figure 1.1.1: Flow chart of Total Carbon (Sepa Chemistry<sup>1</sup>, ES-INR-P-004)

Moxley (2011), states that the concentration of organic carbon in many Scottish rivers, has approximately doubled over the last twenty years, with soils being the most likely source. According to Moxley (2011), the rate of TOC increase, averaged across all sites with increasing concentrations, was 0.12 milligrams per litre per year (mg/l/y), giving an increase in TOC concentration of nearly 2.5 mg/l over a twenty year period. However, this increase in organic carbon is not a unique trait of Scottish waters. Increasing DOC concentrations in rivers of boreal and sub-boreal regions have been observed across Great Britain over the past 10 years (Worrall et al., 2007). Furthermore, there was an observed increase in DOC concentrations over large regions in the Northern Hemisphere in the last few decades (Monteith et al., 2007; Weyhenmeyer et al., 2009). Moreover, increases in DOC concentration have been observed in North America (Driscoll et al., 2003; Stoddard et al., 2003), central Europe (Hejzlar et al., 2003), and Scandanavia (Skjelkvale et al., 2001).

Increasing levels of DOC are a cause for concern, as Worrall et al., (2004) explain that the removal of DOC from water resources is a major cost to water treatment in large parts of Britain. If DOC is not removed properly: it can result in water of low aesthetic quality; it can lead to water failing to meet the colour criteria specified in the Drinking Water Directive; it increases the threat of biological contamination of treated water; and can result in the formation of tri-halomethanes, which are potential carcinogens and if present in drinking water, are very dangerous.

## **1.2 Factors Driving Trends**

Many papers which discuss factors driving increasing trends in organic carbon focus on DOC concentrations. The relationship between TOC and DOC shall be explored in Chapter 2 – if there is a strong relationship between the two types of organic carbon, it is plausible that factors thought to be associated with a change in DOC concentrations, will also be associated with a change in TOC concentrations.

Several explanations have to been put forward to explain the recent observed increase in DOC concentrations. Freeman et al., (2001a) have associated observed increases in DOC with the rising temperatures over the previous decades. Worrall et al., (2004) explains that an increase in temperature, leads to greater microbial activity and enhanced decomposition of peat and thus increased production of DOC; but, it is unlikely that temperature increases alone could explain the observed DOC concentrations. However, increases in temperature can lead to a greater draw-down of water tables in the summer, especially when rainfall is low, which in turn increases the depth of the zone of oxidation and production of DOC (Evans et al., 2002). Having said this, Worrall et al., (2004) believe that changes in temperature and water table depths are not the only factors which will have influenced changes in DOC concentrations, other factors will also contribute. Naturally, drought can also cause the water-table depth to drop, and once the water tables have recovered, can trigger anaerobic production of DOC (Worrall et al., 2007). However, as the water tables decline during a drought, Clark et al., (2005) suggest that sulphides low in the peat profile are oxidised, generating high sulphate concentrations that suppress the DOC release; but, as the water table rises, decreasing sulphate inputs could also increase DOC release. The frequency and severity of droughts is thought to be increasing as a result of climate (change) (Sniffer, 2006).

Krug and Frink (1983) proposed that the increasing levels of DOC concentration in the UK could be explained by the decreasing mineral acidity, particularly sulphate (and possibly also nitrate). Sulphate deposition has led to acidification particularly in areas where the soil and geology have limited buffering capacity. The deposition is thought to have suppressed soil microbial activity and so DOC production. As regulation has controlled the emission of sulphate from industrial sources since the 1980s, deposition to land and water has returned to more natural levels and ecosystems have gradually recovered. It is thought that the reduction in sulphate deposition has allowed soil microbial activity to return to more natural levels which has increased DOC production. This argument was supported by Evans and Monteith (2001), as they stated that UK uplands are showing signs of recovering from acidification and suggested that increasing DOC concentrations could be correlated with this behaviour. Whereas acidification of rivers and lochs can be a direct result of a discharge of contaminants— nitrates are often washed into surface waters (from nearby farms) during

heavy rainfall; most sulphate inputs are from the deposition of sulphate released from combustion processes.

Based on studies of DOC concentrations in lakes and streams in Sweden, during the 1970s and 1980s, Tranvik and Jansson (2002) argued that the increase in DOC could be possibly explained by changes in hydrology i.e. a decreasing discharge could result in increasing DOC concentrations. The observed increase in DOC coincided with decreases in temperature and increased precipitation – where increased precipitation during this time period, led to an increased run off from wetland areas and hence an increased DOC flux to these catchments. Alternatively, Worrall et al., (2003) suggested that the increase in DOC could be possibly explained by a change in the flow path through the soil, allowing richer sources of DOC to be accessed.

It is thought that changes to the land management surrounding surface waters (rivers, lakes, streams etc) could possibly explain increases in DOC concentrations. Worrall et al., (2004) explain that afforestation of upland peat, can lead to a significant loss of carbon storage. A disturbance, such as afforestation, will lead to high values of DOC being recorded in the catchment, followed by a sharp decline (Worrall et al., 2004). Upland peat is not only altered by afforestation - draining used to be a popular practice to improve grazing. Drainage makes the water table deeper below the surface (so could be considered a decrease in elevation above sea level or an increase in depth below the surface) and provokes DOC production (due to the increased oxygen supply, as described earlier). Legislation introduced in 1995 prevented further drainage of peat and could possibly explain changes in DOC trends observed after this time period.

## 1.3 SEPA: Sampling and Measuring of Data

The Scottish Environment Protection Agency (SEPA) has provided the data for the purpose of this research. SEPA has provided data on the levels of TOC and DOC concentrations recorded in 333 river sites (time series between: 1983-2010) and 187 loch sites (time series between: 1994-2010) across Scotland; but also provided data for a selection of covariates as discussed in Section 1.3.2. SEPA monitors sites approximately once a month or once a fortnight. Due to independent research projects carried out by SEPA, sampling at some particular sites has been carried out more frequently.

### 1.3.1 SEPA: Measuring DOC and TOC Concentrations

Water samples are generally collected at the river or loch bank, with the exception of a few boat based samples taken from lochs. The sampler reaches out, filling a 100 ml Pyrex glass bottle (to ensure accurate analysis, the glass bottles are free of any organic contaminants), avoiding any local contamination such as dead sheep or detritus. The glass bottles are then taken back to the SEPA lab for analysis. All samples are analyzed within 8 days of being collected. Depending on which region of Scotland, the DOC and TOC concentrations in the sample are measured using one of two methods. However, inter-comparisons tests have been run between the labs, which show that no significant difference is caused by the different techniques. (SEPA Chemistry<sup>1</sup>)

In the SEPA's South East and South West region, TOC concentration levels are determined using chemical oxidation and an *Aurora 1030W TOC Analyser*. The samples are introduced to the reaction vessel of the instrument where the total inorganic carbon (TIC) is removed. Orthophosphoric acid is added to the sample and the acidified sample is sparged with a stream of inert gas as bicarbonates in the sample dissociate to CO<sub>2</sub>. The resulting gas flow is vented for the pre-programmed sparge time. After TIC removal, sodium persulphate

( $\text{Na}_2\text{S}_2\text{O}_8$ ), a strong oxidizer, is added. This oxidant quickly reacts with organic carbon in the sample at 100 °C to form carbon dioxide. When the oxidation reaction is complete, the carbon dioxide is purged from the solution and routed to the NDIR (nondispersive infrared) detector that is sensitive to the specific absorption for the wavelength of carbon dioxide. (Sepa Chemistry<sup>1</sup>)

In the SEPA's North region, thermal oxidation is used. Following acidification and subsequent purging with purified air to remove inorganic carbon (carbonates, bicarbonates, dissolved  $\text{CO}_2$  etc.), samples are injected into a high temperature reactor where, in the presence of a carrier stream of purified air and an oxidation catalyst, elemental and organic carbon are converted to carbon dioxide. The resulting gaseous mixture is swept from the reactor and following cooling, drying and removal of halogen compounds, the  $\text{CO}_2$  content is measured in a *Non-Dispersive InfraRed* (NDIR) detector. The output of the detector is continually monitored by the system's software and the area of the resulting signal peak is converted into a concentration of carbon by comparison with a calibration curve. (SEPA Chemistry<sup>2</sup>)

Dissolved Organic Carbon is measured in the same way as Total Organic Carbon, but samples are filtered (within 3 days of sampling) before analysis to remove any particulate material and leave only dissolved organic carbon (for both thermal and chemical oxidation). SEPA now use a 0.45  $\mu\text{m}$  filter for all samples, but in the past the South East region used a 1.2  $\mu\text{m}$  filter. This will have led to some differences in what is being measured in different locations, but generally the particulate material only makes up a small proportion of the TOC, so the effect of the different filter pore sizes is likely to be small in most cases - it may be more important for more silty samples, for example, after heavy rainfall.

### 1.3.2 Covariates

For each of the sites, SEPA recorded the DOC and TOC concentrations, as described earlier; but also, recorded physical and chemical properties of the sites. For the purpose of this thesis, SEPA has provided data for the following covariates: temperature (degrees Celsius), pH, alkalinity, nitrate concentration (mg/l), sulphate concentration (mg/l) and river flow [note: river flow was only recorded at 49 river sites]. Similar to the DOC and TOC, sampling frequency for chemical parameters was on a fortnightly or monthly basis.

Temperature is measured in the field at the time of sampling using either a conventional thermometer or digital thermometer which has been calibrated.

The pH of a solution is defined by the equation:  $\text{pH} = -\log a_{\text{H}}$  where  $a_{\text{H}}$  is the activity of hydrogen ions in the solution expressed in gram-moles/l (pH levels fall into a scale between 0-14). The pH of a sample is measured using an electrochemical probe which is dipped into the sample. The alkalinity of natural or treated waters is usually due to the presence of bicarbonate, carbonate and hydroxide compounds of calcium, magnesium, sodium and potassium. Alkalinity is measured by titrating the sample with acid to an endpoint of pH 4.5 (pH 4.2 for samples with very low level alkalinity) measured against the pH probe. There are two different types of apparatus which have been used by SEPA to measure the levels of pH and alkalinity: the radiometer *TITRALAB TIM 900*; and the *Metrohm Autotitrator*. However, QC and inter-lab proficiency schemes do not indicate any significant differences between the two forms of measurement (SEPA chemistry <sup>3</sup>, SEPA chemistry <sup>4</sup>).

The term ‘loading’ is often used when discussing river chemistry – it refers to the amount of TOC or DOC passing a given point on the river in a given time for a given volume (1 litre). For example, if the concentration is 2 mg/l, a small stream with a flow of 1 m<sup>3</sup>/s will have a loading of 2×1×1000mg/s. The factor 1000 is the conversion of m<sup>3</sup> to 1 litre.

SEPA measure nitrate and sulphate levels in the lab (separately), using an instrument known as the ‘Konelab 30 Analyser’. Nitrate levels are measured based on the methods for the examination of waters and associated materials, outlined in ‘Oxidised Nitrogen in

Waters'(1981); similarly, sulphate levels are measured based on the methods for the examination of waters and associated materials, outlined in 'Sulphate in Waters, Effluents And Solids' (2<sup>nd</sup> Edition, 1988).

Based on the data available for covariates, the relationship between each of the covariates and the organic carbons can be explored; but, also, the covariates data can be used to possibly explain the different DOC and TOC trends and patterns present at each of the sites.

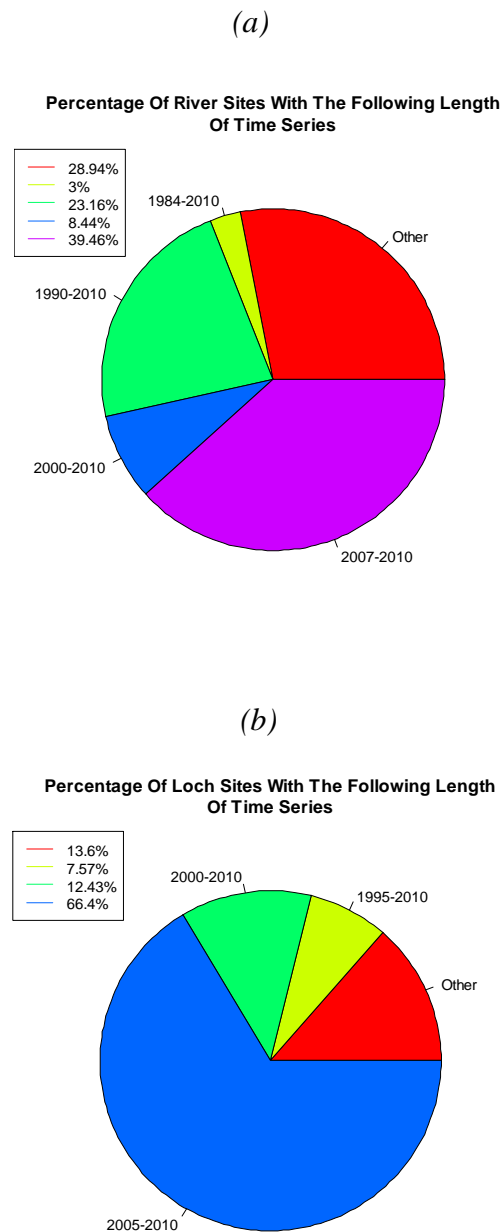
## **1.4 Missing Data and Time Series**

Missing SEPA data can be due to numerous factors: poor weather conditions; staff absences; instrument and analytical difficulties; or revisions to the monitoring plan. Remote sites can be difficult or dangerous to reach during heavy rainfall; or impossible to sample, if frozen over during the winter. The original aim of the thesis was to consider the trends of DOC and TOC across Scotland; but, after exploring the data available for both organic carbons, it became apparent that there were large portions of missing data across the years, with regards to DOC and TOC – distinctly more so for the DOC data. Hence, the decision was made, and agreed with SEPA, to focus on TOC. Data imputation was discussed, but the decision was made to work with the TOC data available – missing data does not present a problem for the standard regression techniques which shall be used later in this thesis for analysis. For both river and loch sites, missing values were present in the following covariates: pH, alkalinity, temperature, nitrate and sulphate. Due to the seasonal behaviour of temperature, the missing temperature values were imputed (as discussed in Section 2.7) and used in the analysis; however, the missing values of the other covariates (pH, alkalinity, nitrate and sulphate) were not imputed.

There was a wide variability in the lengths of time series. Hence, the most common lengths of time series are summarised in Figure 1.4.1. The pie charts highlight the increase in the number of sites added to the study in the past five years – the highest percentage of time series falling into this category for both rivers and lochs. There are only 3% of the river sites



with a time series greater than 20 years – the longest being between 1984 and 2010; and only 7.57% of loch sites with a time series greater than 10 years – the longest being between 1995 and 2010. (Note, “Other”, refers to those sites which had a time period which did not fall into the main categories noted, and generally, sites with fairly small time series).



**Figure 1.4.1: Summary of length of time series at river (a) and loch (b) sites with regards to the TOC data available.**

## 1.5 Overview of Thesis

This thesis aims to:

- Carry out a detailed investigation into the trends of total organic carbon in Scottish rivers and Scottish lochs.
- Investigate environmental and physical factors which could possibly be driving any of the observed trends of total organic carbon in Scottish rivers and Scottish lochs.
- Measure the coherency of the total organic carbon levels between different sites located in Scottish regions.
- Find a model which suitably explains the behaviour of total organic carbon across rivers and lochs in Scottish regions.

To gain an initial impression, Chapter 2 explores the trends and seasonal patterns of TOC across Scotland, but also, investigates the relationship between TOC and the different covariates. The relationship between TOC and DOC is also explored graphically and formally.

Following from this, Chapters 3 to 6 explore the behaviour of TOC in a more formal manner – the thesis aims to move from capturing the behaviour of TOC at one single site, to understanding the trends, and what factors are driving these trends, across Scotland. Chapter 3 simply starts with considering a selection of individual sites and explores the use of parametric and non-parametric regression techniques, to suitably model the TOC. Moving on from this, Chapter 4 (in detail) considers the trend of TOC over time and space in the River Dee network. Unlike Chapter 3, more than one site is being considered. Hence, Chapter 4 discusses finding a model which suitably captures the trends over time and space, which takes into account the location of sites and their relationship to one another (with regards to

river flow and distance). The River Dee network is still being considered in Chapter 5; however, after finding a suitable model in Chapter 4, Chapter 5 aims to measure the coherency between the sites and distinguish common trends within the time series. Chapter 5 then considers measuring coherency on a larger scale than the River Dee network - the coherency between river and lochs sites (separately), in a selection of Scottish regions is explored in detail. Having investigated the coherency between the river and loch sites (separately) in Chapter 5, Chapter 6 aims to build a model which appropriately captures the behaviour of log TOC in each region, taking into consideration time and space. Finally, Chapter 7 ends with a summary and discussion of the findings within the thesis and also discusses possible future work.

# Chapter 2

## Exploring Trends, Seasonality and Relationships

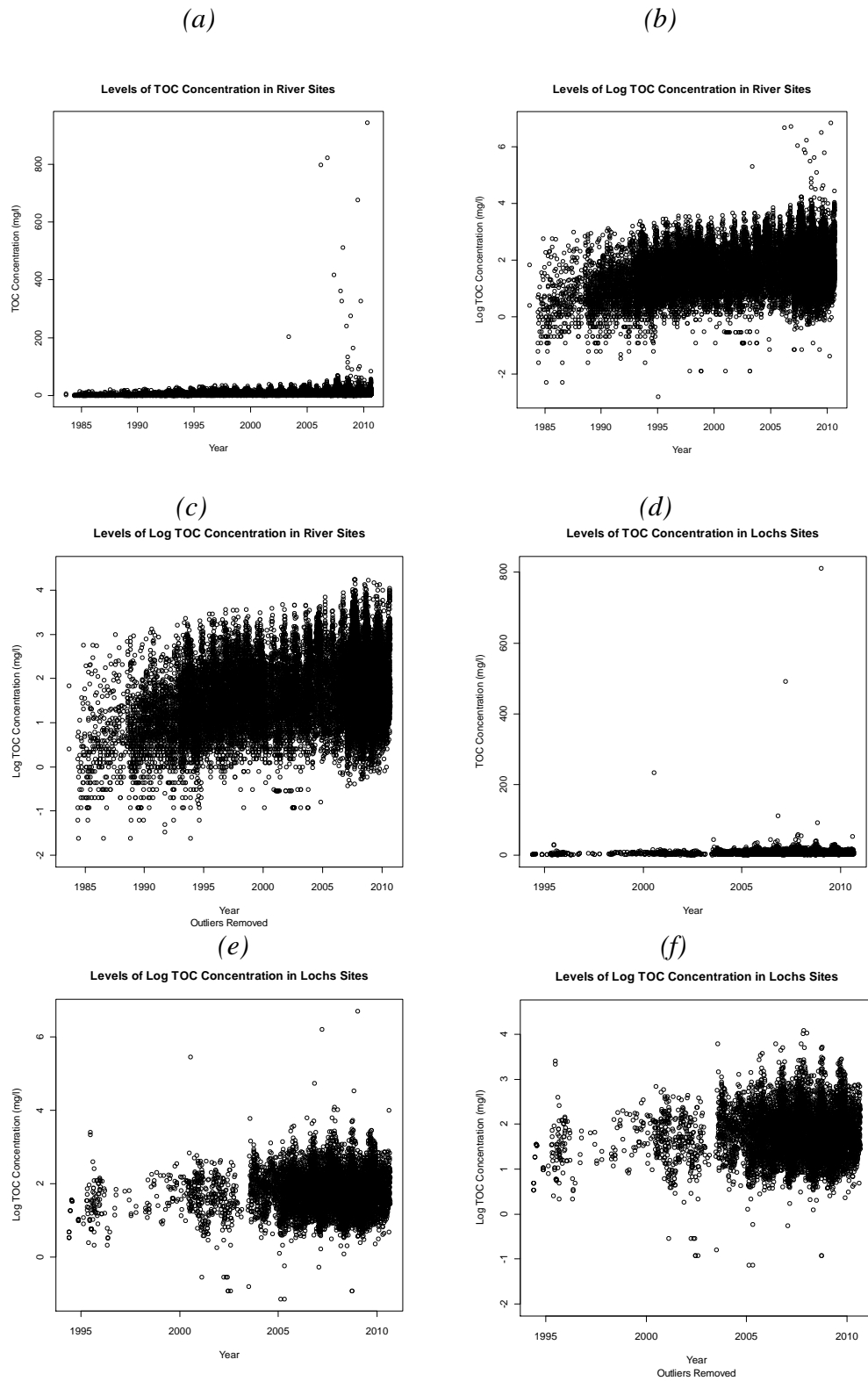
In this chapter, graphical tools shall be used to gain an initial impression of the distribution of Total Organic Carbon at river and loch sites across Scotland. This chapter explores different transformations of the data to find a suitable way of stabilizing the variability in TOC levels across the years. Having found an appropriate transformation of TOC values, scatter plots, box plots and descriptive statistics are used to investigate the trends and seasonal patterns. After exploring the trends and seasonal patterns, this chapter will then switch its focus to obtaining an understanding of the relationships between the transformed TOC values and different covariates. The covariates under investigation are: temperature (degrees Celsius), alkalinity, pH, flow ( $\text{m}^3\text{s}^{-1}$ ) [only data available for 49 sites], nitrate (mg/l) and sulphate (mg/l). Before exploring the relationship between the transformed TOC values and the different covariates, certain issues regarding the covariates had to be addressed. Similar to TOC, the distribution of the data (for each covariate) was examined and different transformations performed. Also, issues regarding values at the limit of detection, with the covariates nitrate and sulphate, will be discussed in this chapter, as well as suitably imputing missing temperature values.

## 2.1 Initial Impression of Total Organic Carbon

To gain an initial impression of the distribution of TOC in rivers and lochs, the levels of TOC from all sites (for rivers and lochs separately) were simply plotted against time, as Figure 2.1.1 [(a) and (d)] displays. Figure 2.1.1 [(a) and (d)] highlights the non-constant variability in the data over time, but also, highlights possible outliers in the data. Therefore, in order to stabilize the variability, different transformations of the data were explored. It was found that taking the log of each value seemed to be the most appropriate for both river and lochs, as Figure 2.1.1 [(b) and (e)] displays, respectively. However, it is clear that even after the log transformation, outliers were still present. Hence, values that were deemed (based on visual exploration of the data) to be outliers were removed from the data set – 0.14% of TOC values were removed from the rivers and 0.06% of the TOC values were removed from the lochs. Having removed the outliers, the log TOC was plotted against time again, as seen in Figure 2.1.1 [(c) and (f)] for rivers and lochs, respectively.

Inspecting Figure 2.1.1, allows an insight into the trends of log TOC present in rivers and lochs. Considering Figure 2.1.1 (c), the plot suggests that the levels of log TOC in the river sites seems to steadily increase from 1985 until the early 2000's, after which the values start to "level off". The plot effectively highlights the wide variability at the beginning of the time period and between 2007 and 2010. The latter is most likely due to the increase in number of sites being monitored by SEPA.

Figure 2.1.1 (f) emphasizes the missing data for the loch sites between 1993 and early 2000; however, from the data available, it appears that the lochs follow a similar pattern to river sites with regard to log TOC (although, the pattern is a little weaker and there seems to be less variability in the values).



**Figure 2.1.1: Scatter plots of TOC against year at river (a) and lochs sites (d); Log TOC against year at river (b) and loch (e) sites; Log TOC against year with outliers removed at river(c) and loch (f) sites.**

The river sites seem to have a lower level of log TOC than lochs on average – rivers having a mean value of 1.68 mg/l compared to the 1.79 mg/l of lochs. This is not unexpected as lochs are generally located in areas with peatier, higher carbon soil, so likely to have more inputs than rivers. The wider variability of the log TOC values in river sites than loch sites is supported by the quartile ranges expressed in Table 2.1.1. However, it is important to remember, that there are log TOC data from 333 river sites, compared to the 187 loch sites being considered – so that, the difference in variability could be due to the smaller number of lochs.

<b>Summary of Log TOC Levels (mg/l) at the 333 River Sites</b>					
Minimum	1 <sup>st</sup> Quartile	Median	Mean	3 <sup>rd</sup> Quartile	Maximum
-1.89	1.16	1.64	1.68	2.2	4.24
<b>Summary of Log TOC Levels (mg/l) at the 187 Lochs</b>					
Minimum	1 <sup>st</sup> Quartile	Median	Mean	3 <sup>rd</sup> Quartile	Maximum
-1.14	1.41	1.78	1.79	2.13	4.09

**Table 2.1.1: Summary of log TOC levels (mg/l) at river and loch sites**

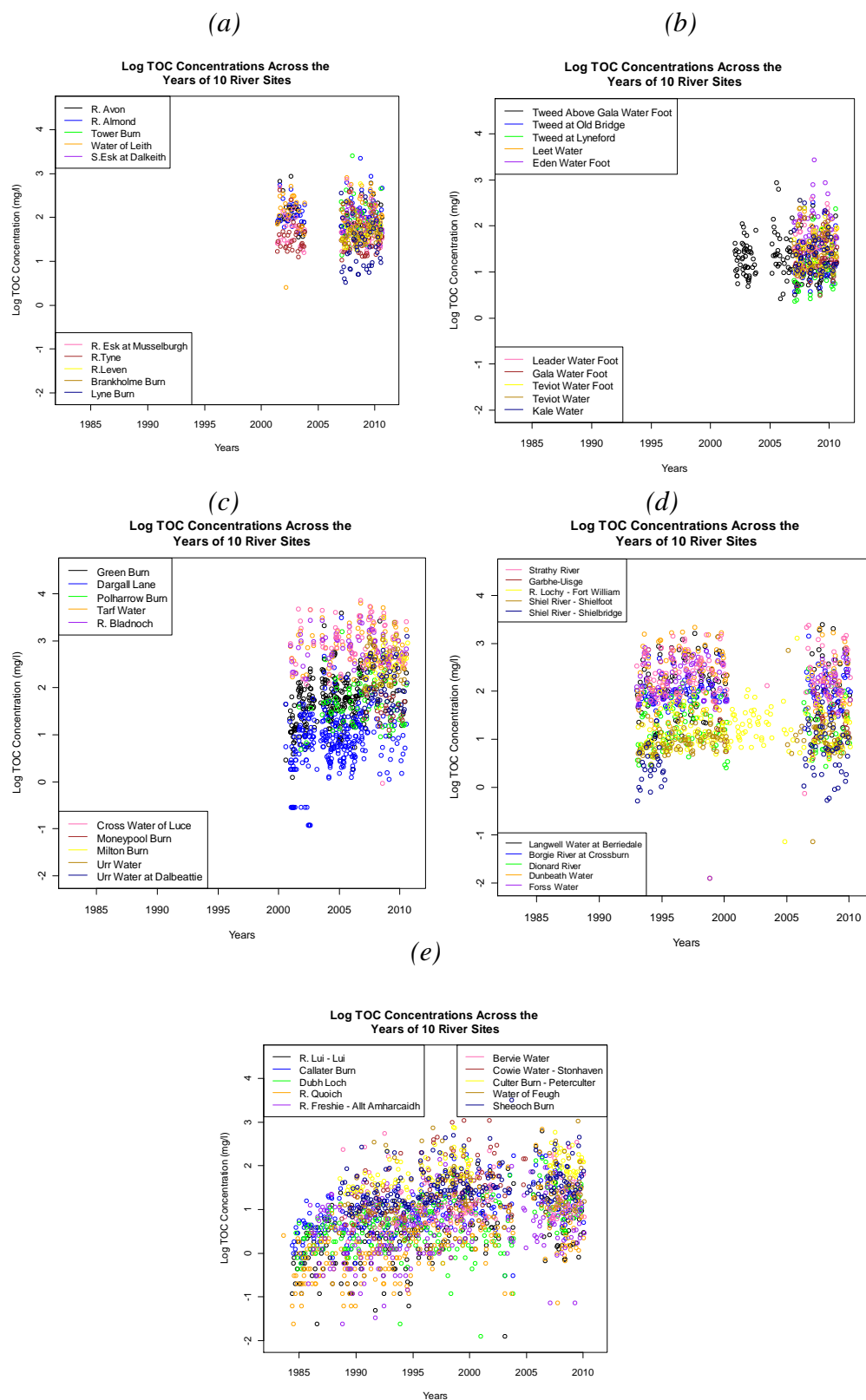
## 2.2 Specific Trends of Log TOC in Rivers and Lochs

In Section 2.1, the levels of TOC and log TOC at the river and loch sites were explored as a whole, providing an indication of the overall trend. However, studying the levels of log TOC at the individual river and loch sites is more beneficial, as it allows the trends of the sites to be looked at in more detail. A selection of the sites have been chosen to represent the most common trends of different river and lochs sites over the different lengths of time series. Figures 2.2.1 and 2.2.3 presents the trends at ten sites on each plot, for rivers and lochs respectively; and Figures 2.2.2 and 2.2.4 look at the trends of a selection of individual river and loch sites in more details. A loess curve (which is based on weighted least squares) has been fitted to the plots in Figure 2.2.2 and 2.2.4 to highlight the presence of any trends.

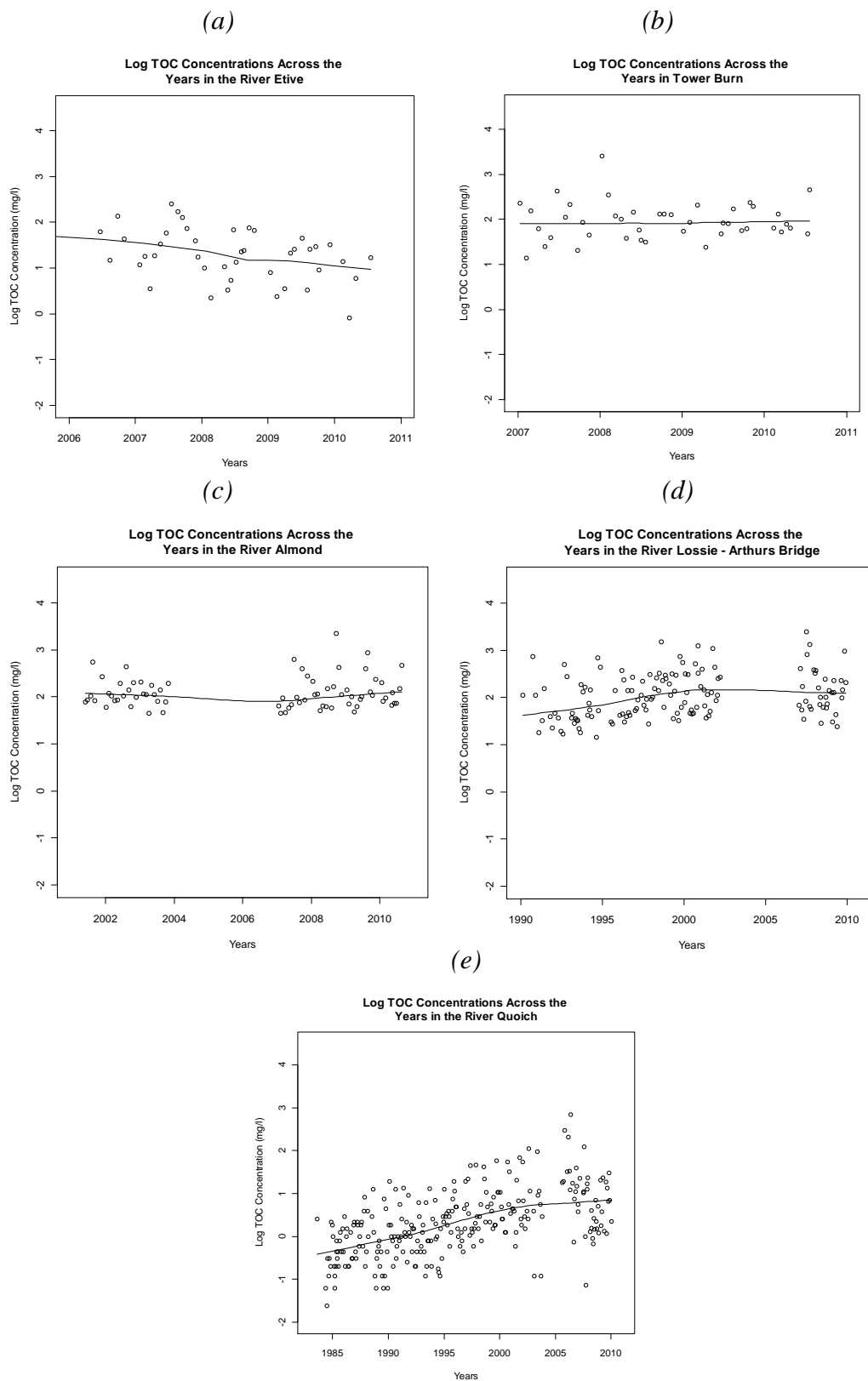
The individual river sites displayed Figure 2.2.1 [(a)-(e)] suggests that the levels of log TOC either remains fairly flat between early 2000 and 2010 or shows signs of a slight decrease. The more detailed plots of the River Eive and Tower burn in Figure 2.2.2 [(a) and (b)] also support this impression. The missing data around 2005 in many of the river sites is also highlighted in Figure 2.2.1 and in the Rivers Almond and Lossie in Figure 2.2.2 [(c) and (d)]. For the river sites with data available from 1983, the level of log TOC seems to steadily increase up until the early 2000's (as shown in detail by the River Quoich in Figure 2.2.2 (e)), after which, the sites follow the same pattern of those with data between the early 2000's and 2010 – the log TOC levels either level off or slightly decrease.

Comparing the river to loch sites, there appears to be similarities between the trends displayed in Figures 2.2.1 and 2.2.3: the level of log TOC in lochs appears to increase from the early 1990's until the early 2000's, which is followed by a levelling off or slight decrease in levels. But, this comparison, again, highlights the wider variability of log TOC levels present in rivers sites, similar to Section 2.1. Figures 2.2.3 [(a)-(e)] and 2.2.4 [(a)-(e)] also stress the missing data present at individual sites – Loch Glashan in Figure 2.2.4 (e) represents a group of loch sites that have large periods of missing data between 1995 to 2000.

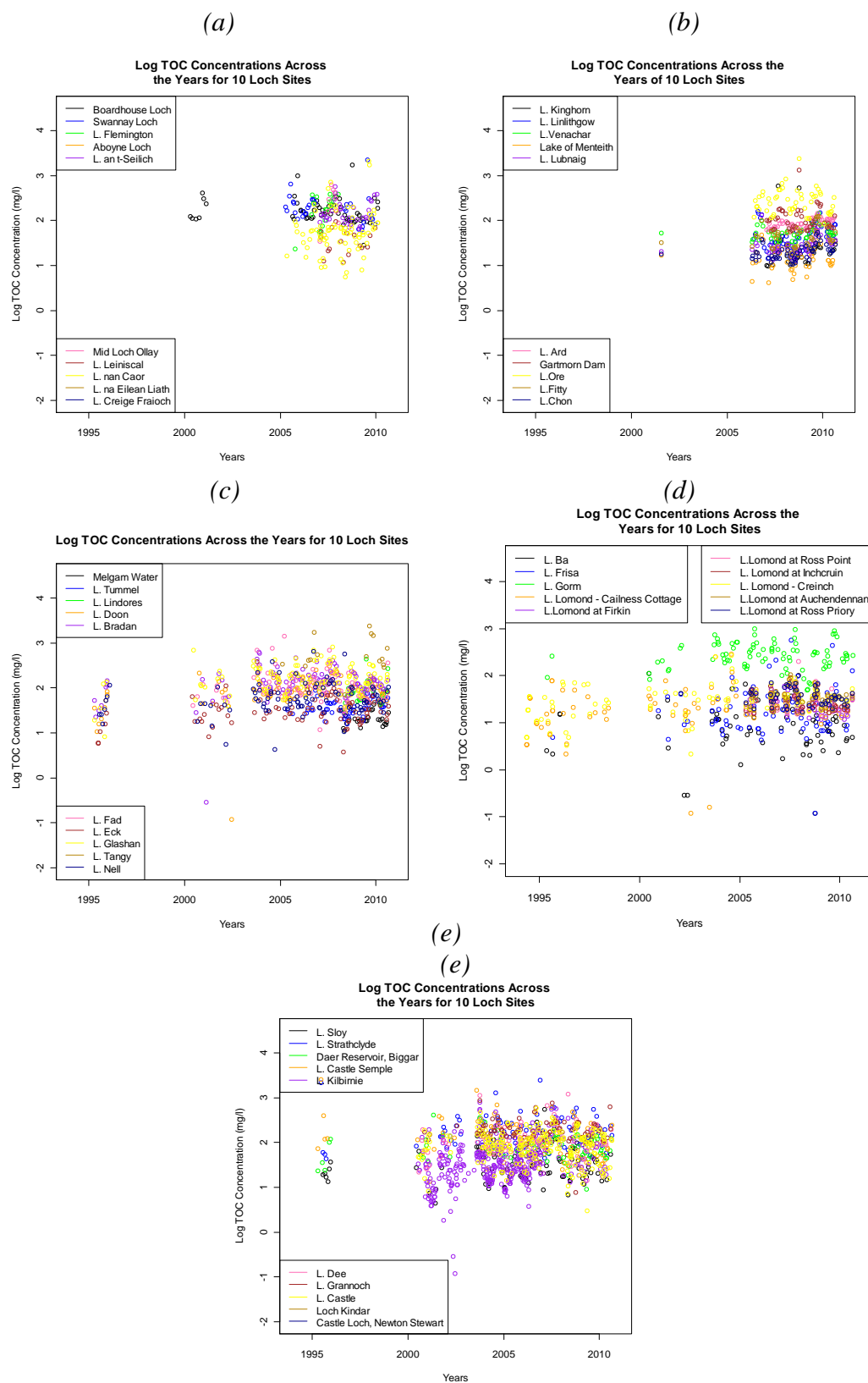




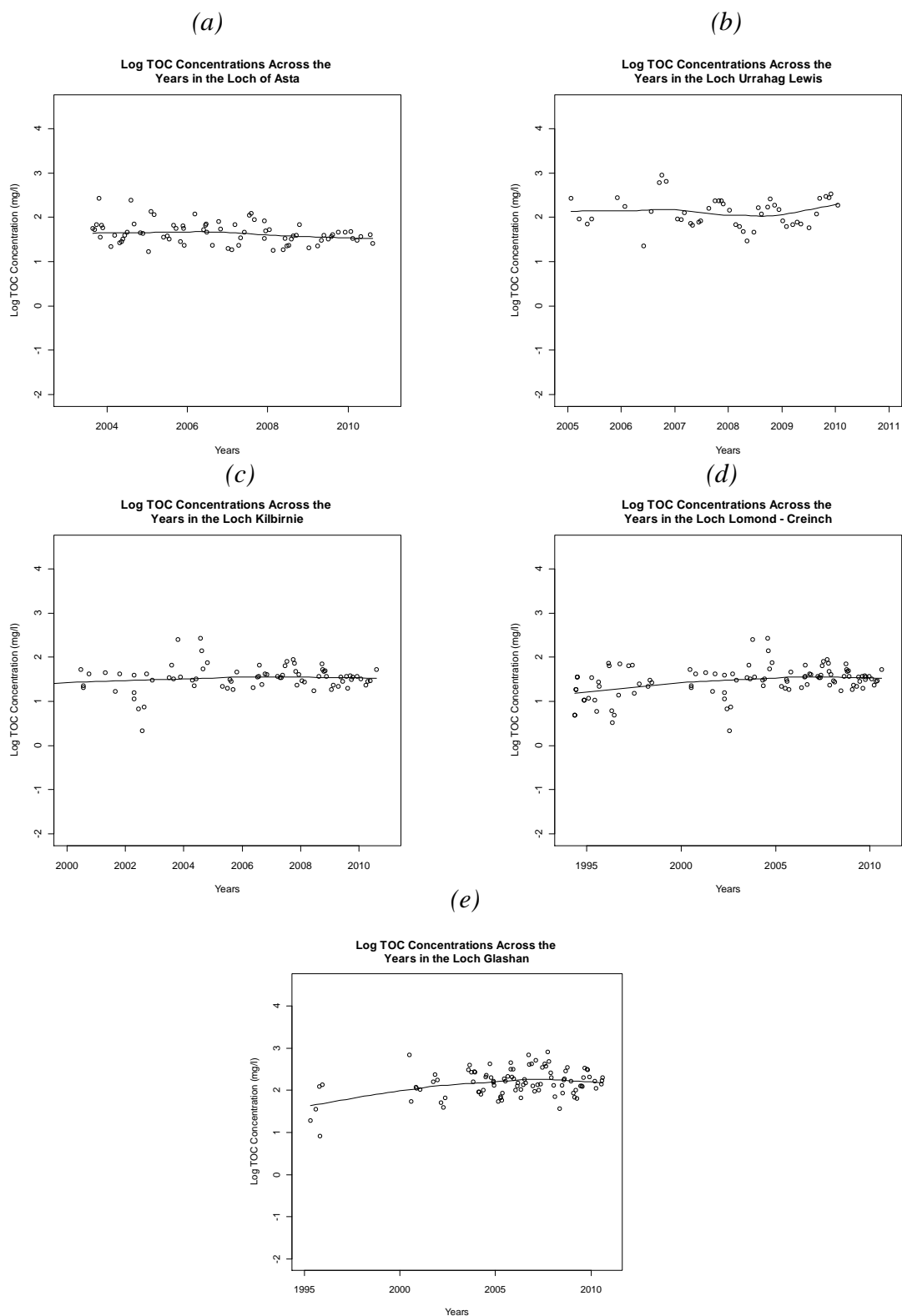
**Figure 2.2.1: Scatter Plots of Log TOC against Year at a selection of river sites, with 10 Sites displayed on each plot.**



**Figure 2.2.2: Scatter Plots of Log TOC against year at a selection of individual river sites with a lowess curve fitted.**



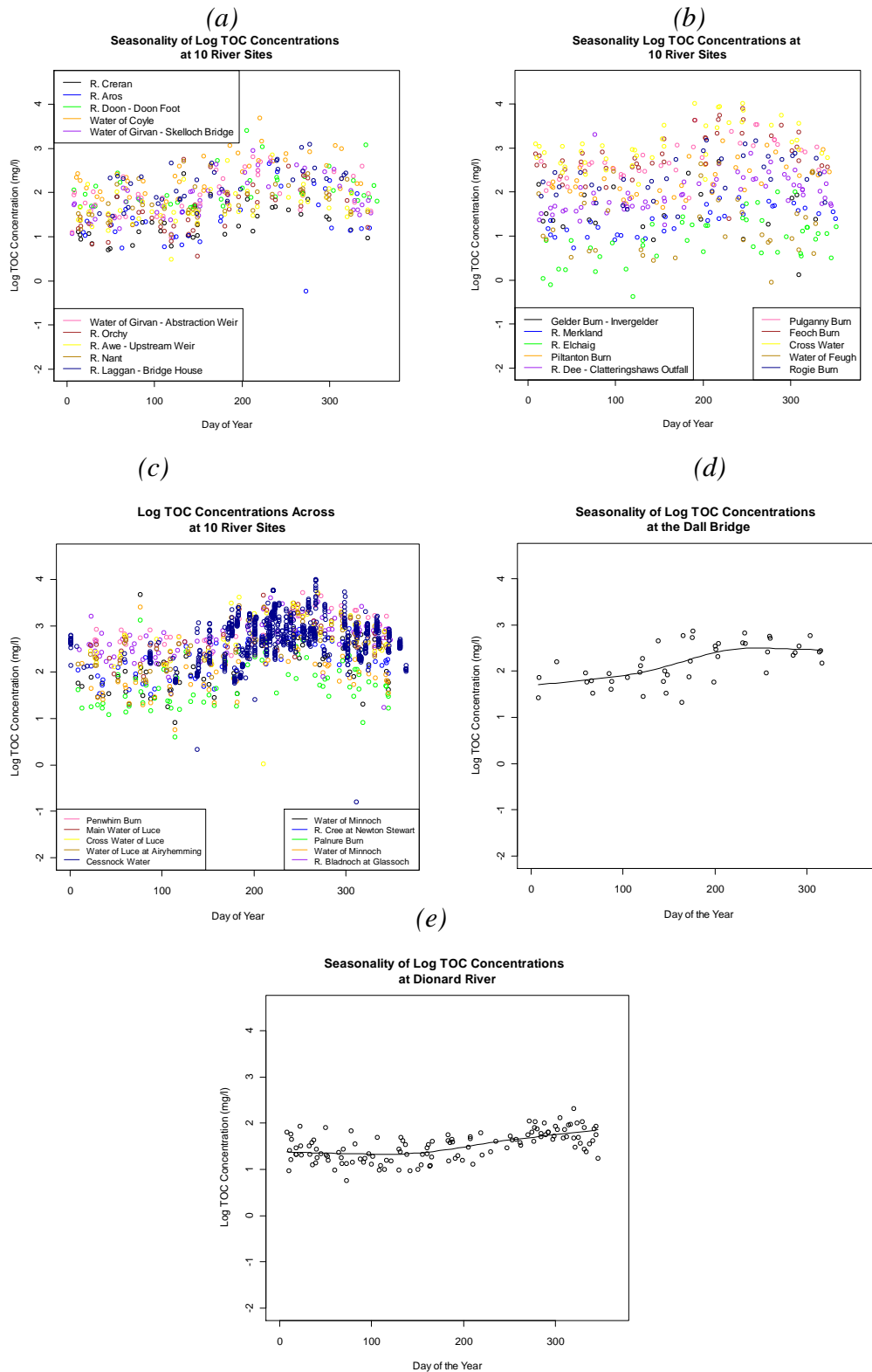
**Figure 2.2.3: Scatter Plots of Log TOC against year for a selection of loch sites. 10 sites displayed on each plot.**



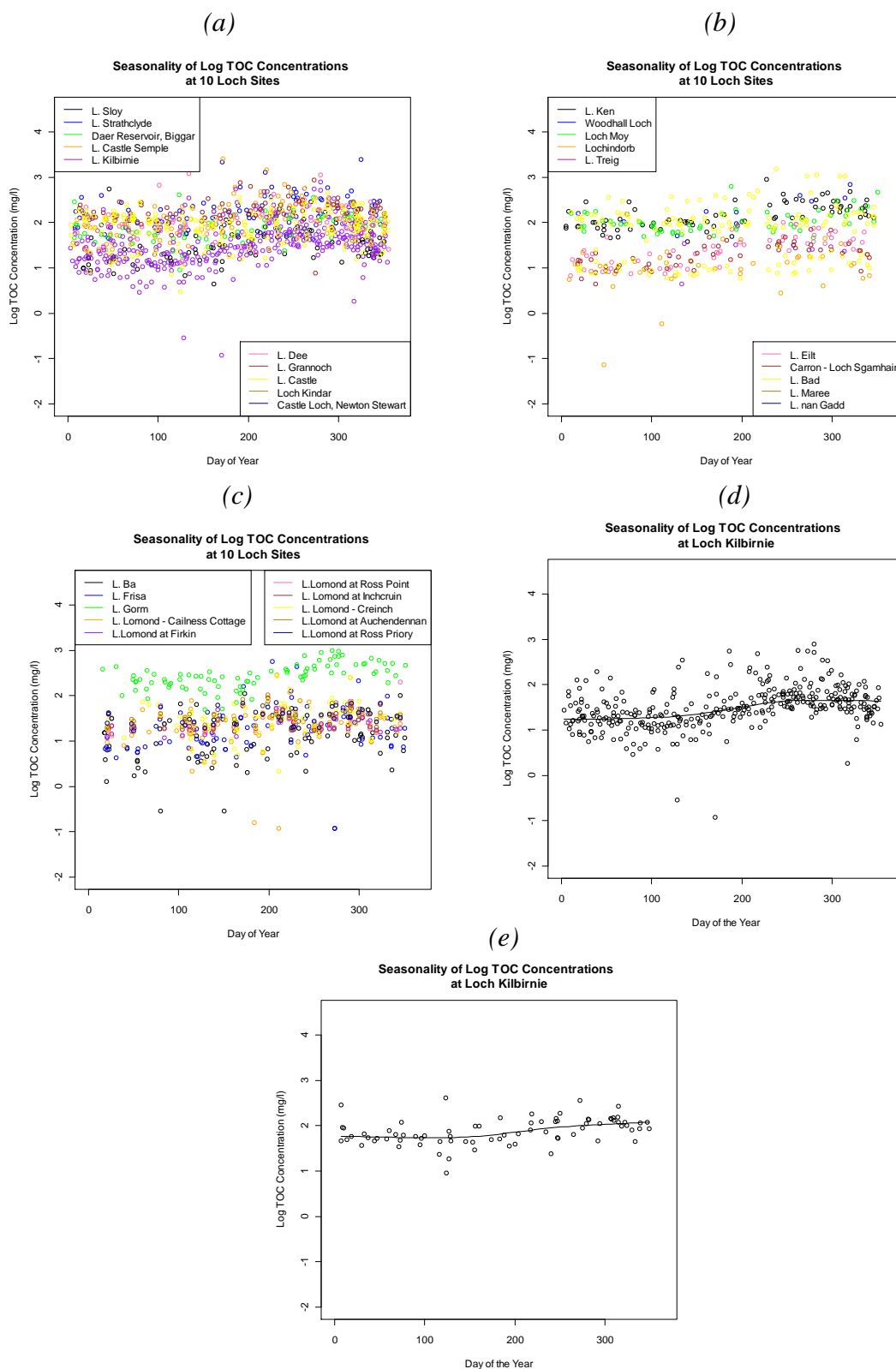
**Figure 2.2.4: Scatter Plots of Log TOC against year for a selection of individual loch sites with a lowess curve fitted.**

## 2.3 Seasonality of Total Organic Carbon in Rivers and Lochs

Section 2.2 explored the presence of trends; but, exploring the presence of any seasonal patterns at the sites is also important. In order to obtain an initial impression of the seasonality, the log TOC levels were plotted against the day of the year in which they were sampled e.g. 1<sup>st</sup> February relates to the 32<sup>nd</sup> day of the year. The seasonality of individual sites was explored at a selection of river and loch sites, as Figures 2.3.1 and 2.3.2 displays – again, a loess curve is fitted to a selection of plots to highlight any seasonal patterns. Considering the plots in Figures 2.3.1 and 2.3.2, the log TOC levels appear to be following a seasonal pattern, in both rivers and loch sites. The levels of log TOC appears to be lowest during early spring, which is followed by a gradual increase until early autumn, when the log TOC levels decline. Log TOC levels seem to be at their highest during the month of September, for both rivers and lochs. Comparing Figure 2.3.1 to Figure 2.3.2, would suggest that the river sites appear to have a stronger seasonal pattern and a wider variability of log TOC levels (in all seasons), than the lochs – similar to the trends discussed previously.



**Figure 2.3.1: Scatter Plots of Log TOC against day of the year for a selection of river sites (a)-(c); and plots of individual river sites with a loess curve fitted (d)-(e).**



**Figure 2.3.2: Scatter Plots of Log TOC against day of the year for a selection of loch sites (a)-(c); and plots of individual river loch with a loess curve fitted (d)-(e).**

## 2.4 Relationship between TOC and DOC

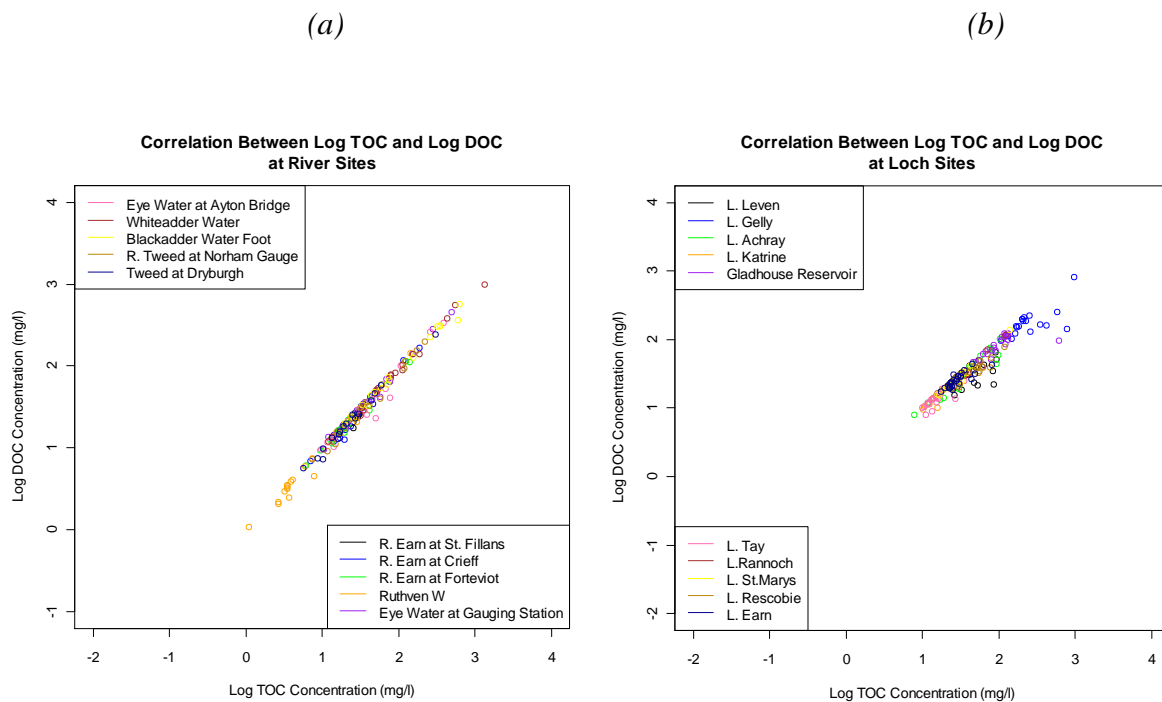
As explained previously, there was a lack of available data for dissolved organic carbon, so the decision was made to only consider the behaviour of TOC. However, a question of interest remains – is there a strong relationship between DOC and TOC? If the two organic carbons are highly correlated, it is possible that factors associated with a change in TOC, would also be associated with a change in DOC.

Thus, to gain a subjective impression of the relationship between the two organic carbons, the TOC and DOC values were simply plotted using a scatter plot. Note, since section 2.1 found the log transformation of TOC to be appropriate, a log transformation of DOC was also performed, and these values used for analysis. When plotting the data, it was essential to overcome some common environmental problems: there was differing amounts of data available for each variable; they were not necessarily collected on the same date; and there was missing data. Hence, the data had to be matched in an appropriate manner to allow an investigation into their relationship. At first, an attempt was made to class a value, sampled on a certain date, as the fortnight of that year in which it fell. For example, the 8<sup>th</sup> January 2005, would be the first fortnight in the year 2005. But, this was not effective, as it failed to provide an adequate number of pairs for analysis. Therefore, the window of matching was increased to a month, so that a sample from the 8<sup>th</sup> January 2005, would be classed as month 1 of the year 2005. If there was 2 samples in a given month, of a given year, an average of the two values would be taken and used for analysis (although two samples in a given month was rare). Note, for the purposes of the rest of this chapter, and the proceeding chapters, the log TOC and covariates samples were classified in this manner, with regards to the date on which they were sampled.

Having found a suitable method for matching the data, a selection of river and loch sites were chosen for analysis. At each of the sites, the log DOC was plotted against the log TOC, as Figure 2.4.1 [(a) and (b)] displays. In both river and lochs, there seems to be a strong relationship between TOC and DOC. The scatter plot suggests that TOC and DOC are highly associated with each other: at river and loch sites, as the level of log TOC increases, the level of log DOC also appears to increase.



Plotting the data provided an informal insight into the relationship. Hence, a formal technique, known as the Spearman's rank correlation coefficient test of association (Best & Roberts, 1975), between the two variables, was implemented. [Note, Pearson's test of association was also considered, but provided very similar results (Hollander & Wolfe, 1973)]. The closer the Spearman's correlation coefficient is to 1, the higher the positive association between the two variables.



**Figure 2.4 1: Scatter plots of log TOC against log DOC at a selection of river (a) and loch (b) sites.**

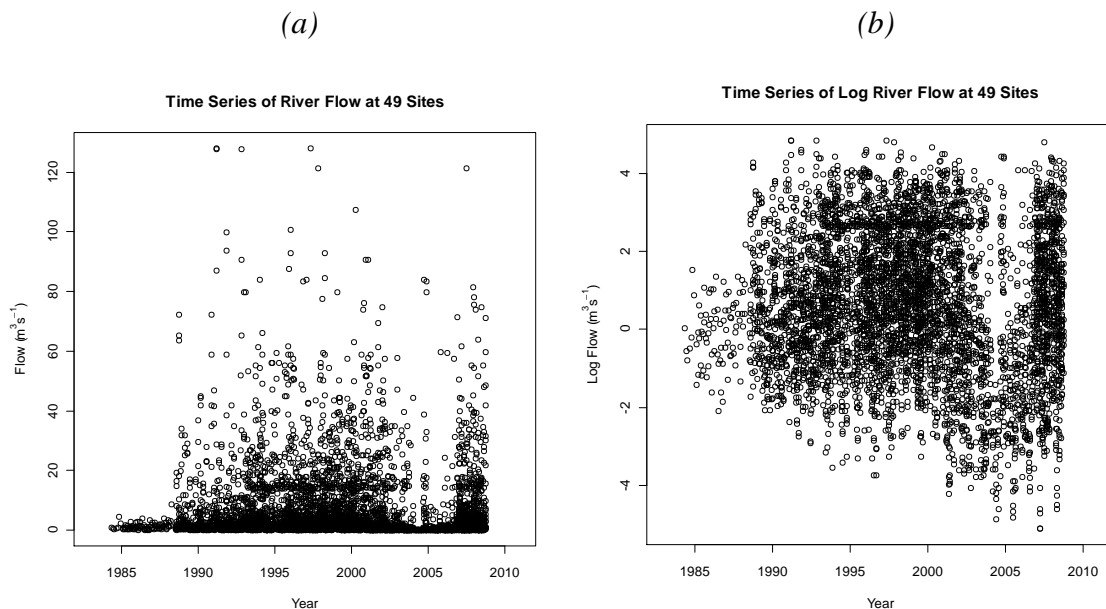
<b>Relationship Between Log TOC and Log DOC</b>			
<b>Rivers</b>		<b>Lochs</b>	
<b>Site</b>	<b>Rho Coefficient</b>	<b>Site</b>	<b>Rho Coefficient</b>
<b>Eye Water at Ayton Bridge</b>	0.91	<b>Leven</b>	<b>0.68</b>
<b>Whiteadder Water</b>	0.99	<b>Gelly</b>	<b>0.68</b>
<b>Blackadder Water Foot</b>	0.99	<b>Achray</b>	0.95
<b>Tweed at Norham Gauge</b>	0.99	<b>Katrine</b>	0.76
<b>Tweed at Dryburgh</b>	0.99	<b>Gladhouse Reservoir</b>	0.88
<b>Earn at St.Fillans</b>	0.89	<b>Tay</b>	0.87
<b>Earn at Crieff</b>	0.94	<b>Rannoch</b>	0.96
<b>Earn at Forteviot</b>	0.98	<b>St. Marys</b>	0.94
<b>Ruthven</b>	0.97	<b>Rescobie</b>	0.88
<b>Eye Water Gauging Station</b>	0.98	<b>Earn</b>	0.85

**Table 2.4.1: Spearman's Rho coefficients for the correlation between a selection of river and loch sites.**

Table 2.4.1 shows the coefficients for the sites displayed in Figure 2.4.1. The Spearman's rank correlation coefficients were fairly close to 1, for most of the river and loch sites (with the exception of Loch Gelly and Loch Leven, which still had a reasonably high level of association). The formal and informal testing of the relationship between log TOC and log DOC at this selection of river and loch sites suggests that there is a strong association between the two types of organic carbon across Scotland. Therefore, based on this, it seems plausibly, that the trend and seasonal patterns displayed in TOC may be similar to that of DOC. Furthermore, it is possible that factors found to be plausible drivers of such trends and seasonal patterns may also have a similar effect on DOC. Lower correlations for lochs might suggest that there are additional processes affecting TOC here e.g algal growth and seasonal stratification/turn over in some lochs.

## 2.5 Further Exploratory Analysis – Log TOC Relationships With Covariates

This section shall focus on the covariates, specifically, their relationship with log TOC. Similar to Section 2.1, the data for each of the covariates was plotted over time to attain an idea of the distribution. After plotting, it became evident that there was a wide variability in the alkalinity, sulphate (mg/l), nitrate (mg/l) and flow data for both rivers and lochs [Note, only flow data for rivers]. Hence, similar to Section 2.1, different transformations of the data were taken and it was found, that the log transformation stabilized the variability in the covariates data appropriately. Figure 2.5.1 provides an example of the effective use of the log transformation with regards to the river flow data available for the 49 river sites. Therefore, throughout this thesis, the log transformation of the alkalinity, sulphate, nitrate and flow data shall be used for analysis.



**Figure 2.5.1:** Time series of river flow at 49 sites with (a) and without (b) the use of the log transformation.

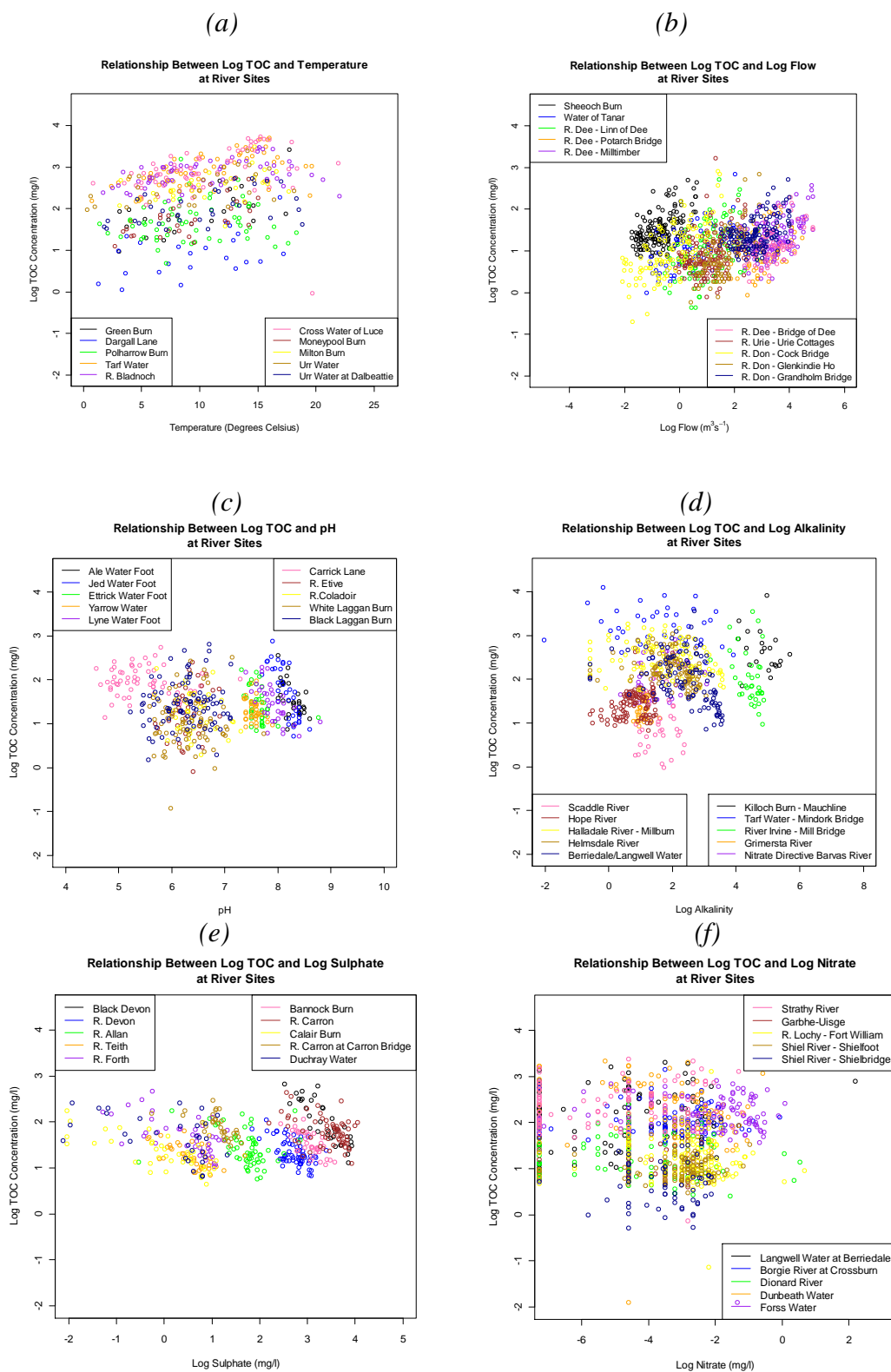
Having inspected the distribution of the covariates, the log TOC of the rivers and lochs could be plotted against each of the covariates to gain an initial impression of their relationships. Figures 2.5.2 [(a)-(f)] and 2.5.3 [(a)-(e)] present log TOC plotted against temperature, log flow (rivers only), pH, log alkalinity, log sulphate and log nitrate at the river and loch sites, respectively.

Comparing Figures 2.5.2 [(a)-(f)] and 2.5.3 [(a)-(e)], allows an insight into whether the physical and chemical effects on log TOC are similar in rivers and lochs. It seems plausible, that the levels of log TOC in both, rivers and lochs, increases as the temperature of the water increases as displayed in Figure 2.5.2 (a) and Figure 2.5.3 (a)- the highest values of log TOC occurring in temperatures of 12-15 degrees Celsius. This suggests that TOC is similar to DOC, in the respect that an increase in temperature leads to great microbial activity, which in turn, increases the production of TOC.

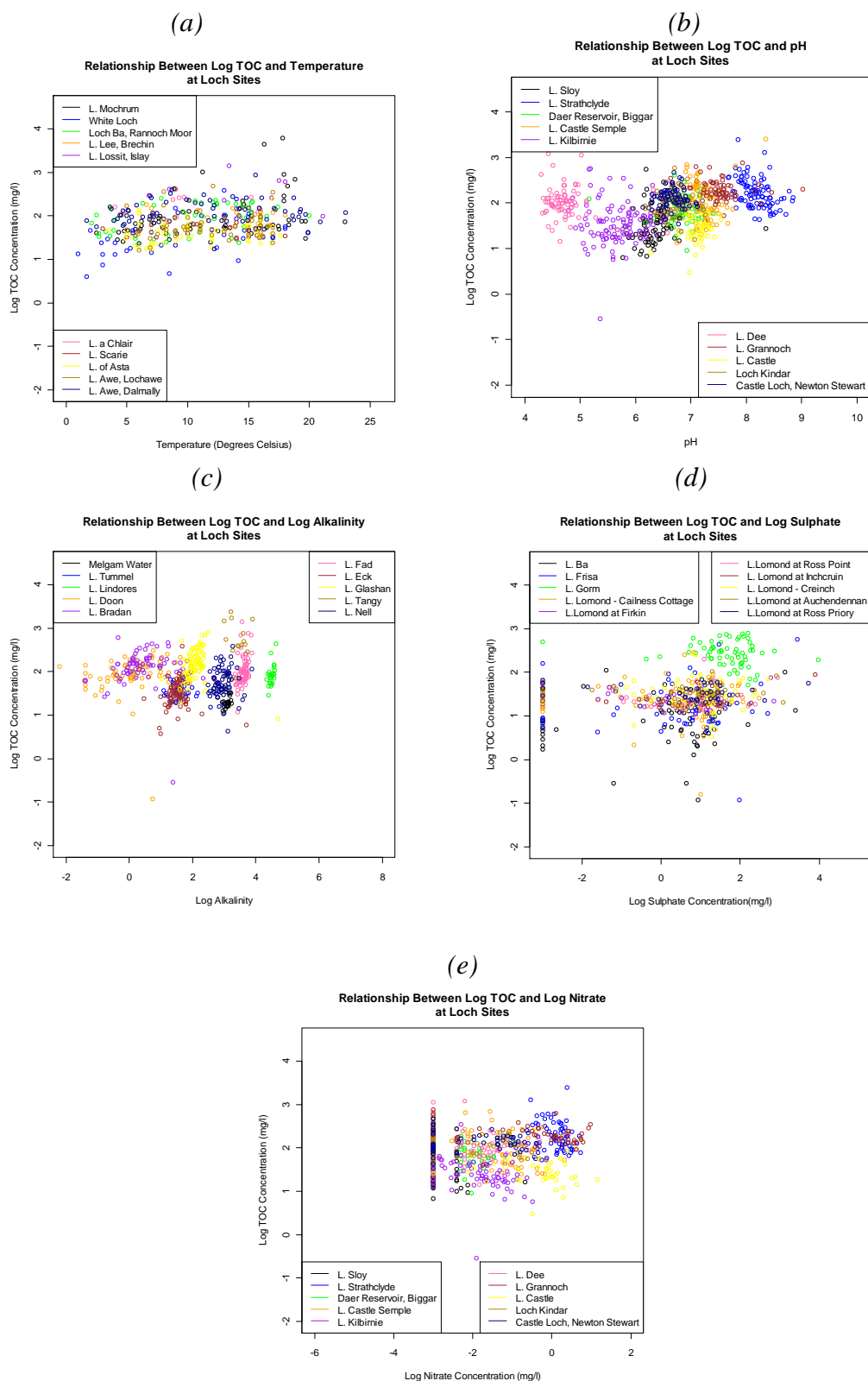
With regards to the effects of pH on log TOC, an overall pattern is not clear. The effects of the variables appear to be site specific in both rivers and lochs. For example, Figure 2.5.2 (c) shows in the River Carrick Lane, that an increase in pH level, is associated with an increase in log TOC - this river is located in the south west of Scotland (Galloway Hills) and so is likely to have been strongly impacted by historical sulphate deposition and acidification. But, this is contrasted by the behaviour observed at Lyne Water Foot [Figure 2.5.2 (c)]: an increase in pH level is associated with a decrease in log TOC levels. In the scatter plots (for lochs) of log TOC against pH, seen in Figure 2.5.3 (b), the points collectively resemble a sine curve, highlighting the site specificity of the effect of pH on log TOC. An increase in pH at some lochs is associated with an increase in log TOC – e.g Castle Loch (similar to the River Carrick Lane, this loch in Dumfries and Galloway is likely to have been strongly impacted by historical sulphate deposition and acidification); but, at other loch sites, it is the contrary, as Loch Strathclyde displays (however, Loch Strathclyde is an artificial water body which was created in the 1970s on an old mining site and is likely to have experienced a variety of pressures different from those in natural and remote locations).

Considering the plot of log TOC against log alkalinity in Figure 2.5.2 (d), it highlights that the effects seem to be differ with each river site. An increase in log alkalinity at Hope River

is associated with an increase in log TOC; however, as seen in Tarf Water – Mindork Bridge, an increase in log alkalinity is associated with the log TOC levels dropping. In contrast, observing the behaviour in the lochs, an increase in log alkalinity seems to be associated with an increase in log TOC at each of the sites, as seen in Figure 2.5.3 (c).



**Figure 2.5.2:** Scatter plots of log TOC against temperature (a), log flow (b), pH (c), log alkalinity (d), log sulphate (e) and log nitrate (f) at a selection of river sites.



**Figure 2.5.3: Scatter plots of log TOC against temperature (a), pH (b), log alkalinity (c), log sulphate (d) and log nitrate (e) at a selection of loch sites.**

The coloured points used in Figures 2.5.2 [(c) and (d)] and 2.5.3 [(b) and (c)], highlight that the pH and log alkalinity levels of each site do not seem to have a wide variability.

Subjectively, considering Figure 2.5.2 [(e) and (f)] and Figure 2.5.3 [(d) and (e)], it does not seem likely that the levels of log nitrate or log sulphate influence the levels of log TOC in either river or loch sites. The log TOC levels remain fairly flat, regardless of any increase or decrease in the log nitrate or sulphate concentration of the water.

As possibly expected, Figure 2.5.2 (b) demonstrates that an increase in the river flow is associated with an increase in log TOC. An increased river flow is generally due to heavy rainfall, which can cause organic carbon to be washed into the water from the soil in the catchment.

## **2.6 Values at the limit of detection: Regression on Order Statistics (ROS)**

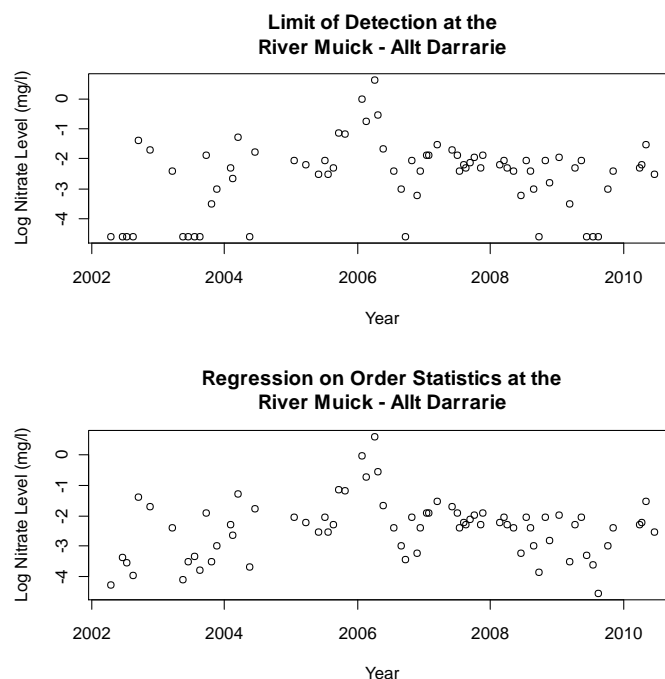
It is clear from Figures 2.5.2 (f) and 2.5.3 [(d) and (e)], that the plots of log TOC against log nitrate in the rivers and log TOC against log nitrate and log sulphate in the lochs, that there appears to be a distinct vertical line of points on the left hand side of the plots. These points are known as values at the limit of detection and are described as left-censored observations. This was not an issue that affected either log DOC or log TOC. Apparatus used to measure any element, has a minimum level which it is able to detect - this is known as the Limit of Detection. In the cases in which a value is at the limit of detection – SEPA halves the value recorded, since it is believed that the true value will lie somewhere between zero and this value. If there were only a few values (generally less than 10%) at this limit of detection, common practice would be to ignore the issue. However, that is not the case, with log nitrate in rivers or log nitrate and log sulphate in lochs. Hence, a method called regression on order statistics (ROS) can be used (Helsel, 2005). It is a semi-parametric method for computing summary statistics of a distribution where there is left censored or non-detect data. Left censored observations are modeled using a linear regression model of the observed un-censored values against their normal quantiles. The ROS method requires the same key assumptions as linear



regression: that the response is a linear function of the explanatory variable or variables and that the variance is constant. However, since it is extremely common in environmental contexts that the variables of interest are skewed, a log transformation of the data is often taken prior to application of the ROS method. (Helsel, 2005).

Figure 2.6.1 shows an example of this regression on order statistics technique in practice. Figure 2.6.1 shows the log nitrate levels in the River Muick – Allt Darrarie with the values clearly at the limit of detection; and the log nitrate levels after performing Helsel’s method.

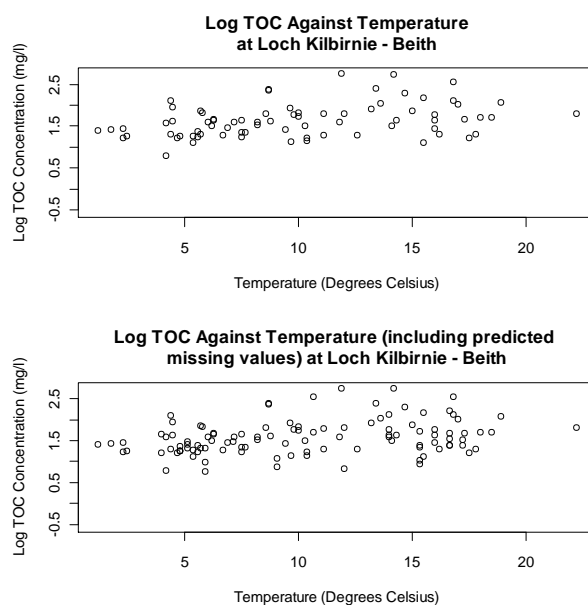
From visual inspection, it seems that the technique is effective and statistically sound. From this point on, the log nitrate levels in the rivers, and the log nitrate and log sulphate levels in the lochs, shall be computed using ROS if they are deemed to be at the limit of the detection.



**Figure 2.6.1: Time series of log nitrate (mg/l) with and without the ROS computation at the River Muick – Allt Darrarie.**

## 2.7 Predicting Temperature

After plotting the log TOC levels against temperature at the river and loch sites, and then having a closer look at the data available for each site, dealing with the missing temperature values became of interest. As temperature tends to follow a seasonal pattern, predicting the missing temperature values could be achieved in a sensible manner. The missing temperature values were simply replaced with the mean temperature value of the month in which it was missing. For example, if the temperature value of Loch Kilbirnie was missing in January 2005, then, the temperature of Loch Kilbirnie in January 2005 would be predicted as being the mean of all the observed January temperatures over the time series. Figure 2.7.1 displays the temperature values over the years at Loch Kilbirnie, with and without the predicted values. Figure 2.7.1 highlights that the predicted temperature values at Loch Kilbirnie seem to fit in with the rest of the data effectively. This method shall be performed on all the missing temperature values at all river and loch sites.



**Figure 2.7.1: Time series of log TOC (mg/l) against temperature (degrees Celsius) with and without predicted missing temperature values.**

## 2.8 Conclusions

The aim of Chapter 2 was to explore the data – firstly, to explore the trend and seasonal pattern of TOC, and then to explore the relationships between TOC and DOC, but also, the relationships between TOC and the covariates.

Section 2.1 highlighted that log transformation of the TOC, suitably stabilized the variability observed in the data. Furthermore, the exploratory analysis suggested that the levels of log TOC in Scottish rivers and lochs are behaving in a similar manner to DOC studied in the Northern Hemisphere, North America, central Europe and Scandinavia (as Discussed in Chapter 1) – there has been an observed increase. The levels of log TOC seemed to increase throughout the 1990's in both rivers and lochs; it is not until the early 2000's that the increase seems to weaken.

Plotting the levels of log TOC against the day of the year in which they were sampled, allowed an inspection of any seasonal patterns. The plots revealed a clear seasonal pattern in both, rivers and lochs – the log TOC levels seemed to be increasing from early spring until early autumn. The seasonal pattern appears to be stronger in rivers.

Having explored the trend and seasonal patterns, the relationships between log TOC and log DOC was of interest. The use of scatter plots and correlation tests (Spearman's and Pearson's) suggested that there was a strong relationship between the two types of organic carbon.

An initial impression of the relationships between log TOC and the covariates could be gained through the use of scatter plots. Similar to the ideas discussed in Chapter 1 by Freeman et al., (2001a) and Worrall et al., (2004), the plots suggest that temperature is also associated with an increase in log TOC levels in Scottish rivers and lochs. Given the short time series for lochs, this may reflect seasonal temperature variation rather than long term year on year climate change. Highest levels of log TOC are associated with a temperature of approximately 12-15 degrees Celsius.

However, with regards to the effects of pH on log TOC, the effect seems to be site specific, for both rivers and lochs. An increase in pH at one site is associated with an increase in log TOC; but, at other sites, it is the contrary. The site specificity is similar in rivers, with regards to log alkalinity effects on log TOC; however, at loch sites, an increase in log alkalinity is associated with an increase in log TOC.

Unlike the other covariates, the levels of log nitrate or log sulphate do not seem to influence the levels of log TOC in either river or loch sites. The log TOC levels remain fairly flat, regardless of any increase or decrease in the log nitrate or sulphate concentration in the water.

Based on visual exploration, it seems likely that an increase in the river flow is associated with an increase in log TOC levels at the site.

The plotting of the different covariates raised two issues – values at the limit of detection and missing values. Log nitrate (in rivers and lochs) and log sulphate (in lochs) had values which were recorded at the limit of detection. To overcome this issue, a technique known as regression on order statistics (Helsel, 2005) was implemented, which seemed to effectively deal with the problem. Now, for the second issue – as temperature generally follows a seasonal pattern, the missing values could be predicted in a sensible manner by simple computation based on the monthly mean.

# Chapter 3

## Modelling Trend, Seasonality and Covariates at Sites

The previous chapters have explored the trends and seasonality of log TOC, but also the relationship between log TOC and the different covariates. However, only a subjective impression has been obtained. The focus of this chapter shall be to investigate three river and three lochs sites. Each site (approximately) represents different time periods of data available and were chosen on the basis that they are assumed to be representatives of the full data set. The river sites under study are: Callater Burn (1984-2010); Dall Bridge at Bridge Main Street (2006-2010); River Tweed above Gala Water Foot (2002-2010). The loch sites to be considered are: Loch Kilbirnie – Beith (2000 – 2010); Loch Lomond – Creinch (1994-2010); and Loch Naver (2005 – 2010). Studying sites with differing lengths of time series will provide an insight into whether the length of time period will have an effect on the observed trends; but, also, whether or not it will influence the relationships between log TOC and the different covariates at each of the sites. In this chapter, suitable modeling techniques shall be explored and a final model shall be chosen to appropriately capture the behaviour of log TOC at each of the sites individually.

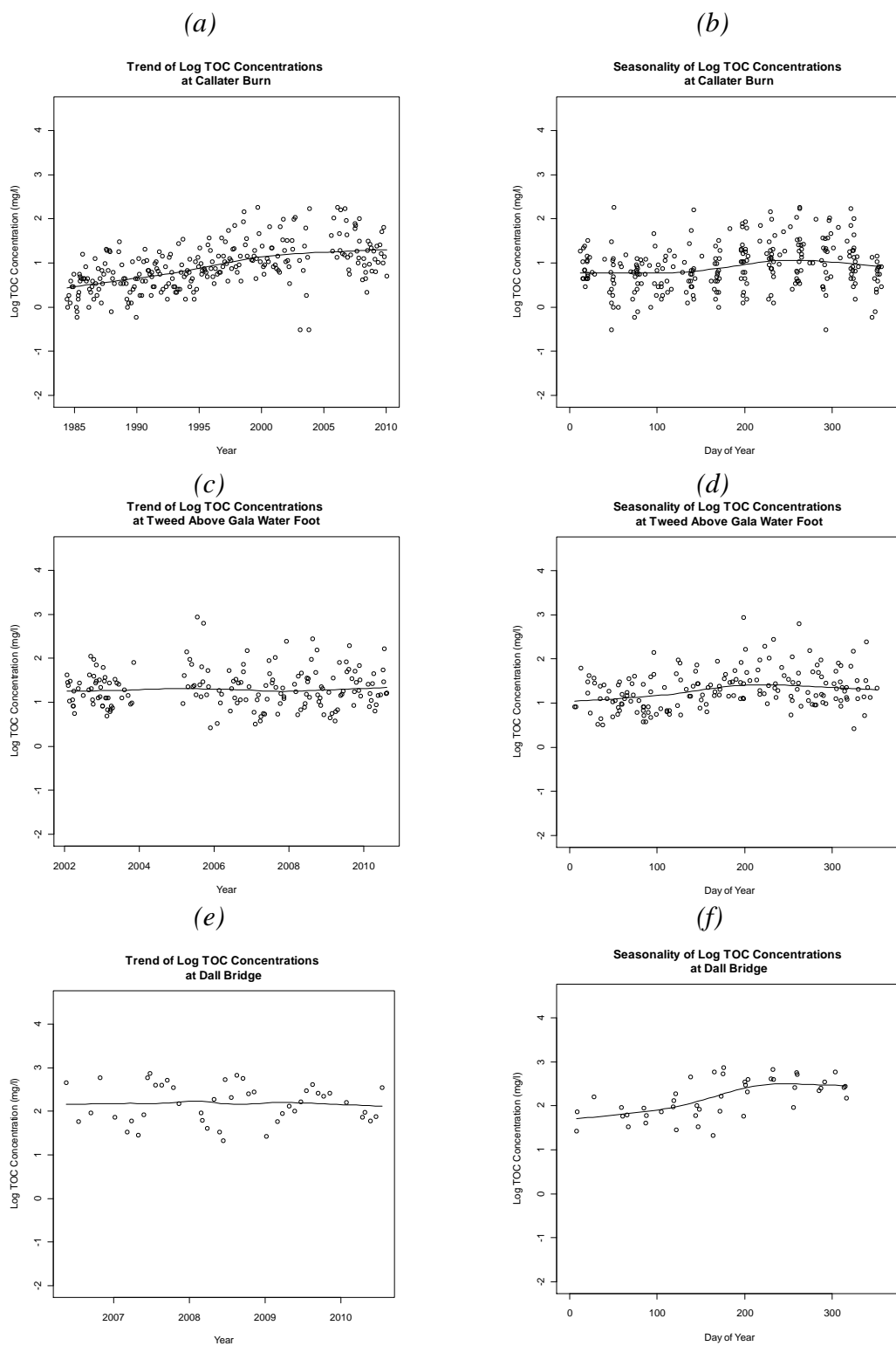
### 3.1 Initial Impression of the sites

The log TOC levels are plotted over time and against the day of the year in which they were sampled, to provide an insight into the trends and seasonal patterns present at each of the sites. Studying Figure 3.1.1 [(a),(c) and (e)], suggests that the trend differs between sites. At Callater Burn [Figure 3.1.1 (a)], there appears to be a consistent increase in log TOC levels from the early 1980's, through until the early 2000's – this constant increase is followed by a “leveling off” of the log TOC levels after the year 2004. This “leveling off” of log TOC levels is apparent in most sites, after the year 2004. The river sites Tweed above Gala Water Foot [Figure 3.1.1 (c)] and Dall Bridge [Figure 3.1.1 (e)], which have data available from a shorter time period than Callater Burn, maintain a constant level of log TOC throughout their time periods.

Switching our focus to Figure 3.1.2 [(a),(c) and (e)], the loch sites show a similar pattern to that of the river sites, with regards to trend. However, the “leveling off” of log TOC levels occurs later. At the Loch Lomond – Creinch site [Figure 3.1.2 (a)], the levels increase from the middle of the 1990's, through until the year 2005. It is from then, that the log TOC levels are fairly constant. This trend is also seen in Loch Kilbirnie (Beith) [Figure 3.1.2 (c)], even with the shorter time period. Loch Naver [Figure 3.1.2 (e)], with five years of data, shows no significant trend, and behaves similar to the other two loch sites in their latter years.

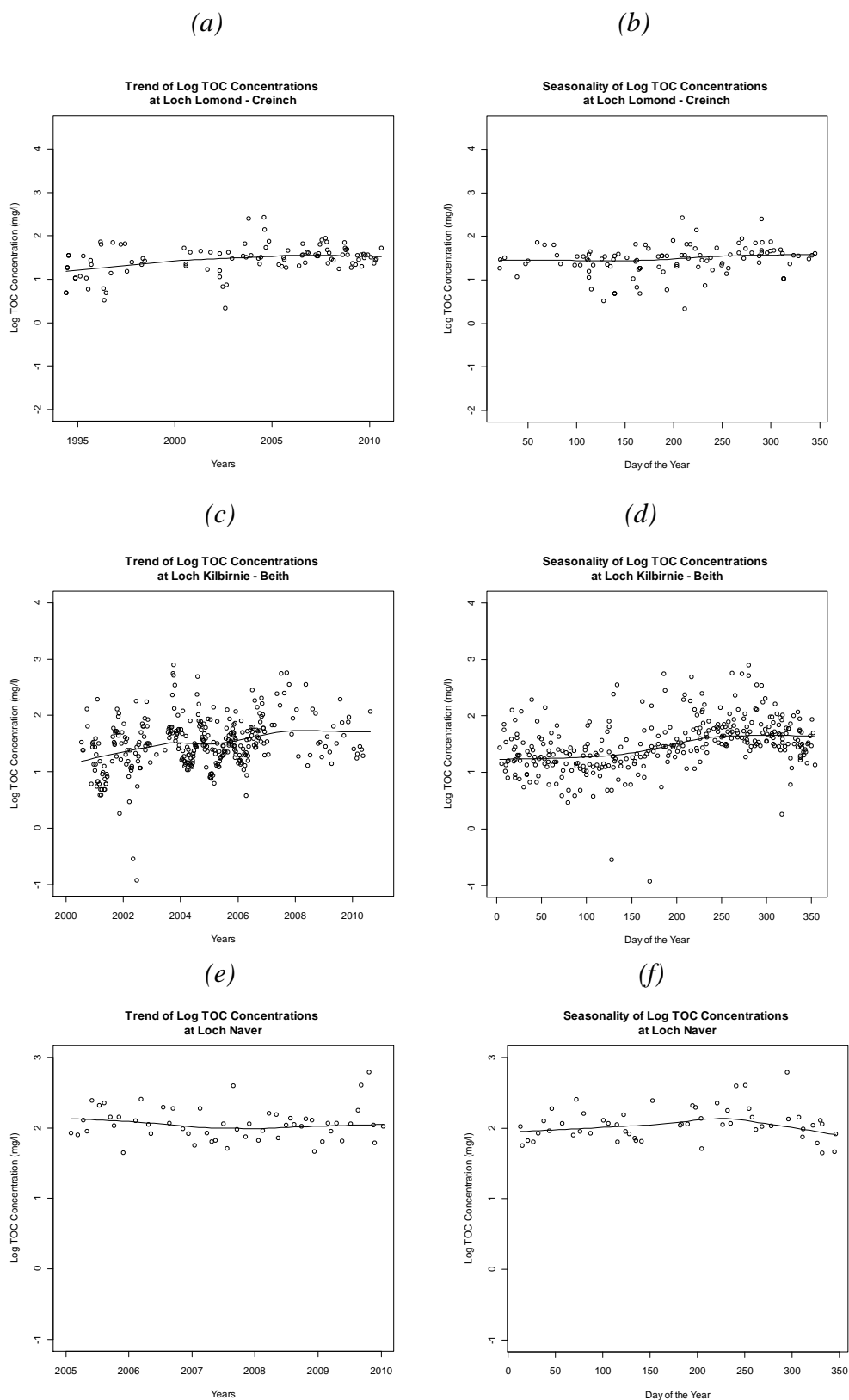
Considering the seasonality of river sites in Figure 3.1.1 [(b),(d) and (f)], it seems fair to say, that the log TOC levels behave in a similar manner. The lowess curve fitted to the plots, suggest that levels are lowest during late winter and early spring and gradually increase throughout the summer until early August. This pattern is more evident at Callater Burn and Dall Bridge [Figure 3.1.1 (b) and (f)], than Tweed above Gala Water Foot [Figure 3.1.2 (d)] – Callater Burn and Dall Bridge are peatier catchments (more soil carbon) which plausibly explains the observed patterns. A visual inspection of the plots in Figure 3.1.2 [(b),(d) and (f)], suggests that the seasonal pattern in the loch sites is not as strong – with the exception of

Loch Kilbirnie [Figure 3.1.2 (*d*)]. The seasonality of Loch Lomond and Loch Naver appears to be rather flat [Figure 3.1.2 (*b*) and (*f*)].



**Figure 3.1.1: Time series of log TOC (mg/l) at the three river sites [(a),(c) and (e)]; and seasonality plots of the three river sites [(b),(d) and (e)] with regards to log TOC levels.**





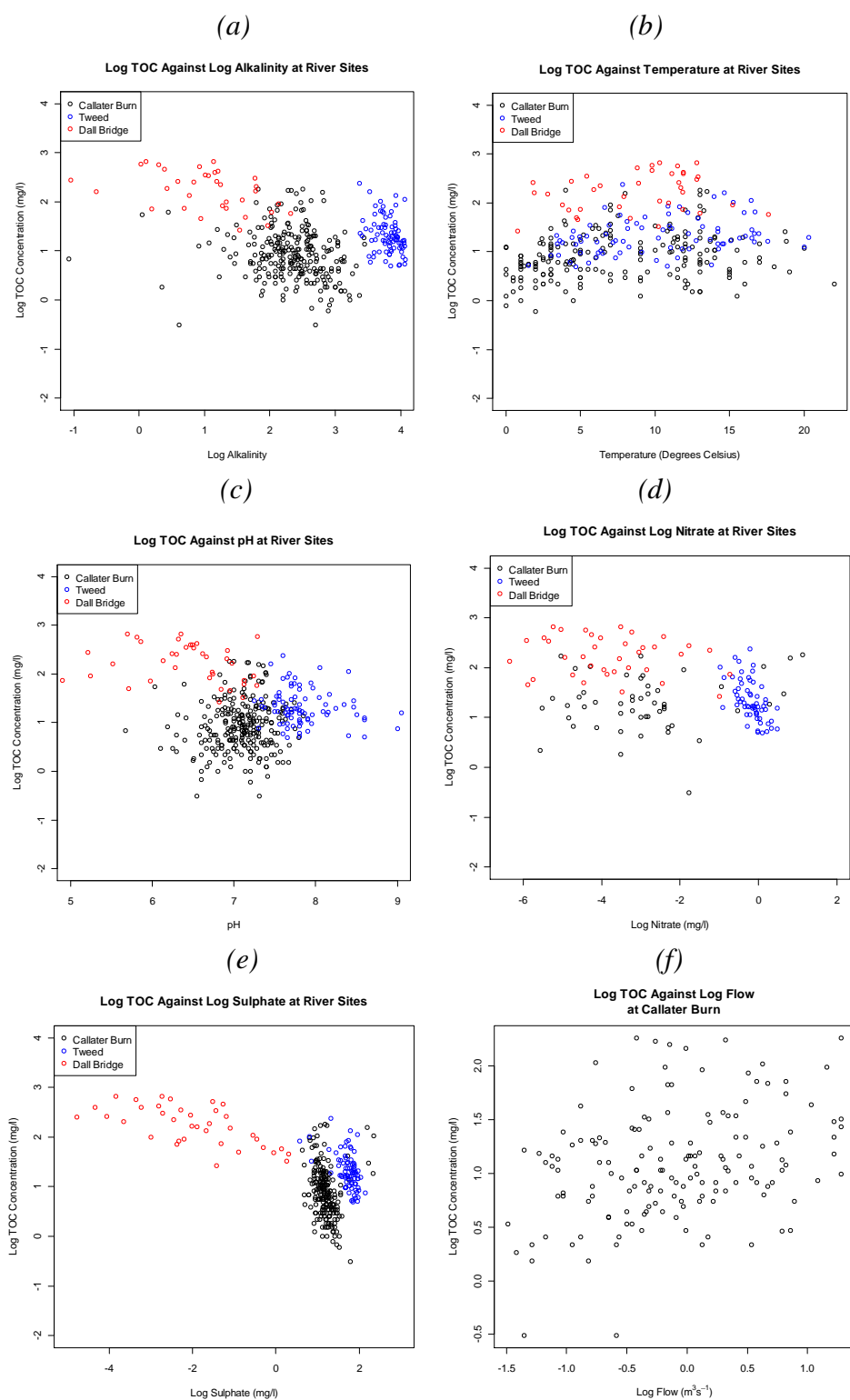
**Figure 3.1.2: Time series of log TOC (mg/l) at the three loch sites [(a),(c) and (e)]; and seasonality plots of the three loch sites [(b),(d) and (e)] with regards to log TOC levels.**

## 3.2 Relationship Between Log TOC and Covariates

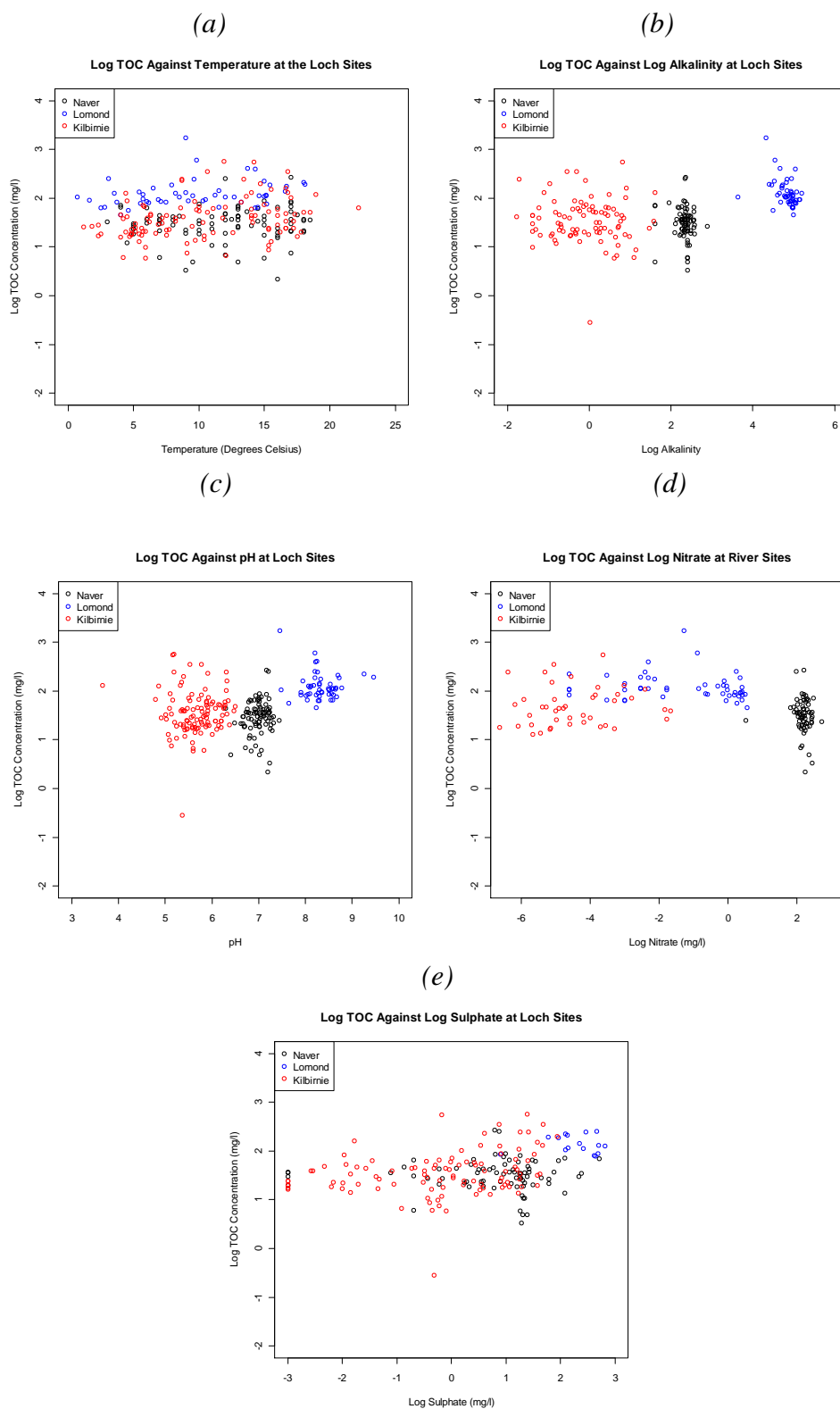
Similar to the previous sections, the relationship between log TOC and the covariates shall be explored through the aid of graphical tools. Figure 3.2.1 [(a)-(f)] displays covariates which appeared to be associated with a change in log TOC levels at the three river sites. Figure 3.2.1 (a) suggests an increase in the log Alkalinity at Dall Bridge and Callater Burn, is associated with a decrease in the log TOC levels. Figure 3.2.1 (b) suggests that an increase in the temperature is associated with an increase in log TOC at all three sites. Figure 3.2.1 (d) suggests that log nitrate appeared to be associated with a decrease in log TOC levels only at the River Tweed. Figure 3.2.1 (e) suggests that an increase in log sulphate is associated with a decrease in log TOC levels at Dall Bridge; but not associated with a change in log TOC levels at the other two sites. Figure 3.2.1 (f) suggests that an increase in log flow at Callater Burn is associated with an increase in log TOC levels.

Figure 3.2.2 [(a)-(e)] leads one to believe, that unlike the river sites discussed previously, the covariates measured at the loch sites do not seem to have any strong relationship with the log TOC. Figure 3.2.2 [(a) and (b), respectively] suggests that an increase in temperature or log alkalinity in Loch Naver, could possibly be associated with a change in log TOC levels – an increase in temperature (optimum temperature, once again, of approximately 10-15 degrees Celsius), could be plausibly associated with an increase in log TOC levels; and an increase in log alkalinity could be associated with a decrease in log TOC. Loch Kilbirnie (Beith) displayed a similar relationship with regards to temperature; but, the other covariates did not appear to show any strong relationships. Considering the relationship between log TOC and the covariates, Figure 3.2.2 [(a)-(e)] reveals no evidence of any strong relationships at Loch Lomond (Creinch).

From comparing Figures 3.2.1 and 3.2.2, it seems plausible that the physical and chemical factors have a different effect on log TOC levels in rivers and lochs.



**Figure 3.2.1: Log TOC plotted against temperature (a), log alkalinity (b), pH (c), log nitrate (d), log sulphate (e) and log flow (f) [Callater Burn only] at the river sites**



**Figure 3.2.2: Log TOC plotted against temperature (a), log alkalinity (b), pH (c), log nitrate (d), log sulphate (e) at each of the loch sites.**

### 3.3 Modelling Log TOC At Each Site

Exploratory analysis is useful in providing a subjective impression of which factors influence log TOC at different sites. A natural progression from exploratory analysis is to move to performing formal analysis at each of the sites. Formal analysis allows us to explore different modelling techniques with an aim of finding a model which appropriately explains the behaviour of log TOC at each site separately.

#### 3.3.1 Harmonic Regression

From the exploratory analysis, it became clear that a trend over time was apparent in those sites with a longer time series; but, a seasonal pattern was evident in all of the sites. Bearing this in mind, a sensible starting point would be to fit a model using harmonic regression.

Harmonic regression is used to incorporate seasonal patterns. For a periodically oscillating observation  $y$  (log TOC), the sine function is used to build a regression model of the form

$$y_i = \alpha + \gamma * \sin\left(\frac{2\pi\{t_i - \theta\}}{p}\right) + \varepsilon_i \quad i=1, \dots, n \quad (3.3.1.1)$$

where  $t_i$  is an independent predictor variable that captures the time effect, (e.g. month),  $\theta$  is an angle of the sine function,  $\varepsilon$  is an additive error term, and the remaining quantities are parameters that affect the nature and shape of the sine wave. If in equation 3.3.1.1 we assume no temporal correlation exists among the error terms, then it may be reasonable to set  $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$  and view the model as relatively straightforward nonlinear regression. If we make the further assumption, that the period  $p$  (e.g. 12 here) is known, then the model can be reduced to simpler multiple linear regression model. That is,

$$\begin{aligned}
y_i &= \alpha + \gamma \sin\left(\frac{2\pi\{t_i - \theta\}}{p}\right) + \varepsilon_i \\
&= \alpha + \gamma \sin\left(\frac{2\pi t_i}{p} - \frac{2\pi\theta}{p}\right) + \varepsilon_i \\
&= \alpha + \gamma \left\{ \sin\left(\frac{2\pi t_i}{p}\right) \cos\left(\frac{2\pi\theta}{p}\right) - \cos\left(\frac{2\pi t_i}{p}\right) \sin\left(\frac{2\pi\theta}{p}\right) \right\} + \varepsilon_i \quad (3.3.1.2)
\end{aligned}$$

the latter equality following from the well known trigonometric ‘double angle’ formula  $\sin(\psi - \phi) = \sin(\psi) \cos(\phi) - \cos(\psi) \sin(\phi)$ . For known  $p$  (months in the year), each of the terms in this expansion can be written in a simpler form. Let  $c_i = \cos(2\pi t_i / p)$  and  $s_i = \sin(2\pi t_i / p)$  be two new (known) regression variables, and take  $\beta_1 = -\gamma \sin(2\pi\theta/p)$  and  $\beta_2 = \gamma \cos(2\pi\theta/p)$  as two new (unknown) regression coefficients. Then, this simplifies to the following multiple linear regression (Piegorisch et al., 2005):

$$y_i = \alpha + \beta_1 c_i + \beta_2 s_i + \varepsilon_i \quad (3.3.1.3)$$

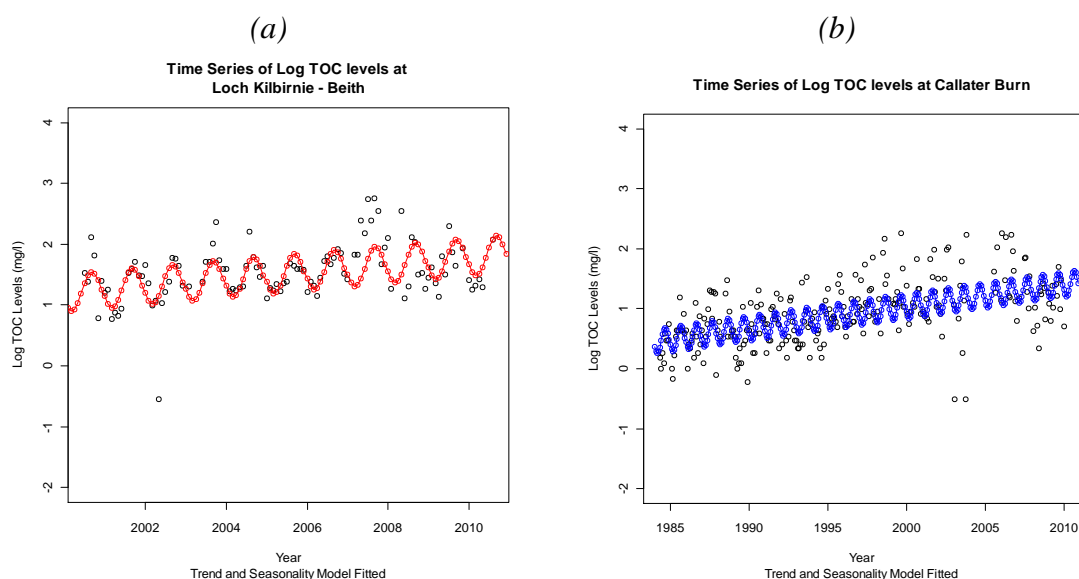
Linear terms, such as ‘Year’ can be easily incorporated into the harmonic regression. For example:

$$y_i = \alpha + \beta_1 Year + \gamma \sin\left(\frac{2\pi\{t_i - \theta\}}{p}\right) + \varepsilon_i \quad (3.3.1.4)$$

Which can also be written as a multiple linear regression:

$$y_i = \alpha + \beta_1 Year_i + \beta_2 c_i + \beta_3 s_i + \varepsilon_i \quad (3.3.1.5)$$

The first step was to consider the trend and seasonality of the sites. Expression (3.3.1.5) was fitted to each of the sites. If the p-value of a fitted term was not significant (i.e. p-value > 0.05), the model was refitted. A summary of the fitted multiple linear regression models can be seen in Table 3.3.1.1. (Note, if either the sine or cosine term in the model has a p-value < 0.05, both terms remain in the model, and deemed to be significant).



**Figure 3.3.1.1: Trend and Seasonality Models fitted to the time series plots at the sites Callater Burn (a) and Loch Kilbirnie - Beith (b).**

Summary of Trend and Seasonality Models Fitted to Sites						
	Sites	Terms	Estimate	St. Error	Pr(> t )	Adjusted R-Sq
Rivers	Callater Burn	Intercept	-72.05	7.09	<0.001	36.5%
		Year	0.04	7.1	<0.001	
		Cos(Month)	0.01	0.01	0.86	
		Sin(Month)	-0.22	0.04	<0.001	
	Dall Bridge	Intercept	2.17	0.05	<0.001	42%
		Cos(Month)	-0.05	0.07	0.5	
		Sin(Month)	-0.37	0.07	<0.001	
	Tweed above Gala Waterfoot	Intercept	1.31	0.04	<0.001	15.1%
		Cos(Month)	-0.07	0.05	0.19	
		Sin(Month)	-0.19	0.05	<0.001	
Lochs	Kilbirnie	Intercept	-116.52	24.01	<0.001	36.17%
		Year	0.06	0.01	<0.001	
		Cos(Month)	0.03	0.05	0.48	
		Sin(Month)	-0.33	0.05	<0.001	
	Lomond	Intercept	-42.27	13.83	0.003	9.4%
		Year	0.02	0.006	0.008	
	Loch Naver	Intercept	2.09	0.04	<0.001	13.07%
		Cos(Month)	-0.04	0.04	0.45	
		Sin(Month)	-0.15	0.05	0.004	

**Table 3.3.1.1: Summary of the final trend and seasonality models fitted to the three river and three loch sites.**

If we focus on the river sites at first, it was found when fitting the multiple linear regression (3.3.1.5), that the trend term (Year) was only significant at Callater Burn (p-value <0.05). Table 3.3.1.1 shows that coefficient for year is 0.04 at Callater Burn. Thus, for any given month, on average, the level of log TOC is increasing by 0.04 mg/l for every one year increase at Callater Burn. The sine and cosine terms fitted in the model were significant at all three sites. Figure 3.3.1.1 (a) effectively shows the use of the harmonic regression – the trend and seasonal model fitted to Callater Burn (highlighted in blue) clearly shows the incorporation of the seasonal effect. However, in saying that, Figure 3.3.1.1 (a), also highlights that there is a lot of unexplained variation, which is confirmed by an adjusted R-Squared value of only 36.5%. This was similar for the models fitted to the Rivers Dall Bridge and Tweed above Gala Water Foot - Table 3.3.1.1 displays the adjusted R-squared values of 42% and 15.1%, respectively.

Fitting expression (3.3.1.5) to the loch sites revealed that the trend was significant at the sites Loch Lomond (Creinch) and Loch Kilbirnie (Beith). Table 3.3.1.1 displays the coefficient for year at each site, respectively, to be 0.02 and 0.06. Thus, for any given month, on average, the level of log TOC is increasing by 0.02 mg/l at Loch Lomond (Creinch) and increasing by 0.06 mg/l at Loch Kilbirnie (Beith), for every one year increase. The seasonal terms were only significant at Loch Kilbirnie (Beith) and Loch Naver as Table 3.3.1.1 displays. Table 3.3.1.1 highlights the amount of unexplained variation in the data from the models fitted to Loch Kilbirnie, Loch Lomond and Loch Naver, with adjusted R-squared values of 36.17%, 9.4% and 13.07%, respectively. The models do not seem to be an adequate fit to the data.



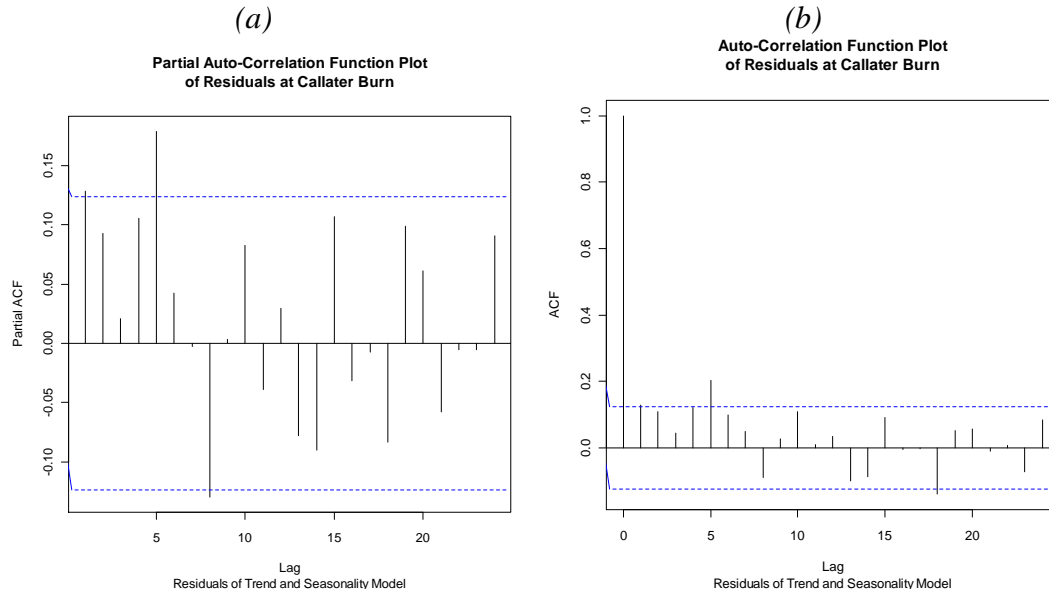
### 3.3.2 Auto-Correlation of Residuals

When modelling the levels of log TOC at river and loch sites, it is plausible that the level of log TOC sampled one month, may be related to the level of log TOC sampled the following month. The correlation between the residuals of linear models can be considered using auto correlation function (acf) and partial auto correlation function (pacf) plots. In the case of a time series, autocorrelation measures the extent of linear relation between values at time points that are a fixed interval (the lag) apart. For a random variable  $X$  at time  $t$ , the population autocorrelation function (ACF) for lag 1,  $\rho_1$ , is given by

$$\rho_1 = \frac{Cov(X_t, X_{t+1})}{Var(X_t)} \quad (3.3.2.1)$$

where the numerator is the autocovariance function for lag 1 and the denominator is the variance of  $X_t$ . At low lags autocorrelation is usually positive. It usually declines towards 0 (for an AR process) as the lag increases. The partial autocorrelation function describes the relation between the lag  $k$  and the corresponding coefficient in an autoregressive model. These plots highlight any patterns or trends of the residuals. (Venables et al., 2002)

Figure 3.3.2.1 [(a) and (b)] displays the ACF and PACF plots of the residuals of the trend and seasonality model fitted to Callater Burn in Figure 3.3.1.1 (a). Figure 3.3.2.1 [(a) and (b)] suggests that there is no significant correlation between the residuals; hence, correlation shall not need to be incorporated in the model. The ACF and PACF plots were produced for the other five sites – similar to Callater Burn, there was no suggestion of significant correlation between the residuals at the sites.



**Figure 3.3.22: Auto-Correlation Function (a) and Partial Auto-Correlation Function (b) plots of the residuals from the trend and seasonality model fitted to Callater Burn.**

### 3.3.3 Fitting Multiple Linear Regression Models

As the correlation of the residuals is not an issue, a natural progression from these preliminary models is to fit a multiple linear regression including the trend and seasonality terms (where appropriate), but also include the other covariates as linear terms. Letting  $y = \log \text{TOC}$ , temperature =  $T$ , log Alkalinity =  $A$ , pH =  $\text{pH}$ , log Nitrate =  $N$ , log Flow =  $F$  (note, flow data only available for Callater Burn) and log Sulphate =  $S$ , the formula for the multiple linear regression can be written as:

$$y_i = \alpha + \beta_1 \text{Year}_i + \beta_2 c_i + \beta_3 s_i + \beta_4 A_i + \beta_5 T_i + \beta_6 \text{pH}_i + \beta_7 N_i + \beta_8 S + \beta_9 \text{Flow} + \varepsilon_i. \quad (3.3.3.1)$$

Where  $y_i$  is the level of log TOC and  $\varepsilon_i$  are assumed to be independent with mean 0 and constant variance. Again, terms that were not significant in the fitted model, were removed, and the model was re-fitted. A summary of the final linear models fitted to each of the sites

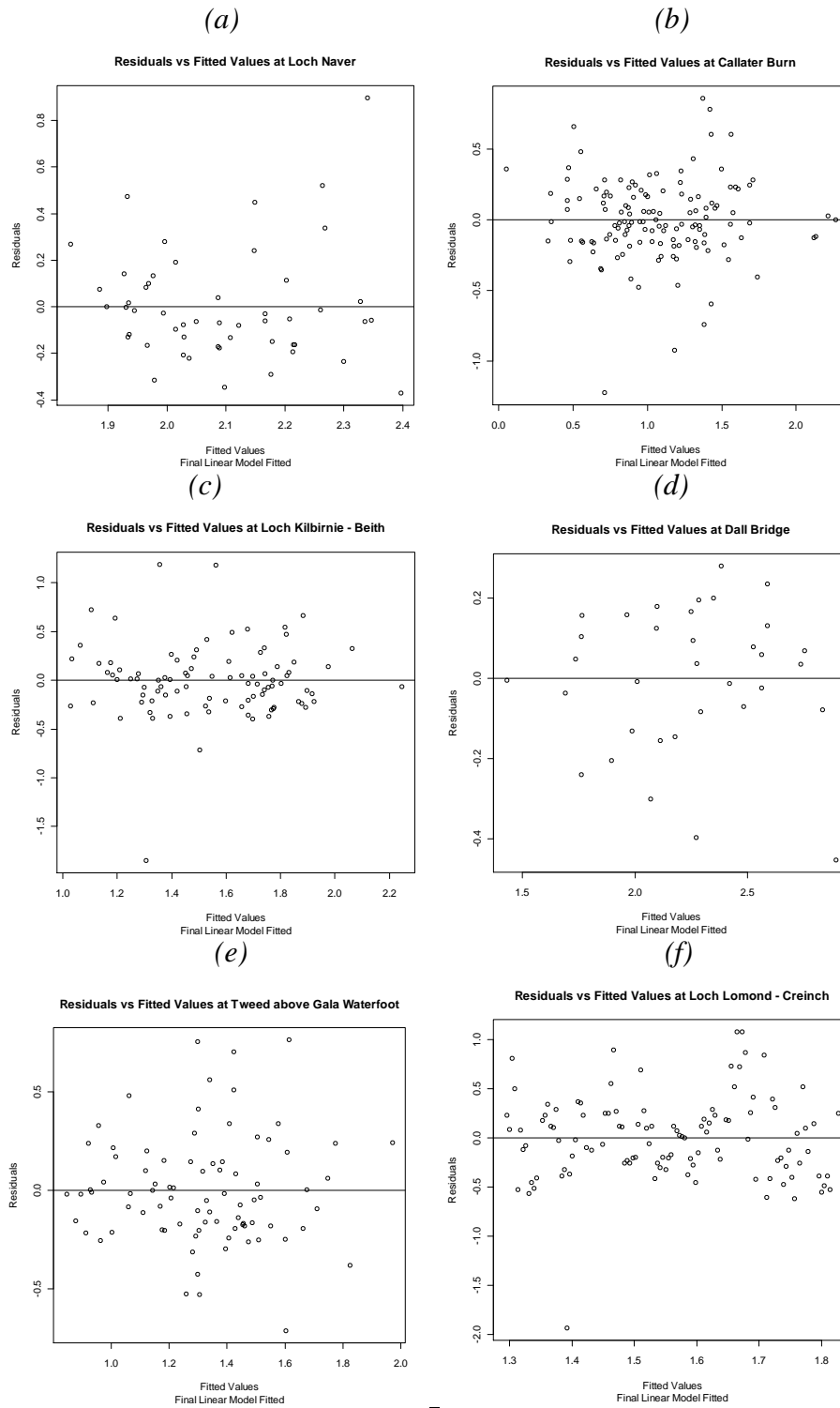
are summarized in Table 3.3.3.1. With regards to the other covariates, only log alkalinity, seemed to have a significant effect on log TOC – and this was only at the sites Dall Bridge, Tweed above Gala Water Foot and the Loch Naver. This suggests that log alkalinity is associated with a change in log TOC levels in both rivers and lochs.

Based on the residuals vs fitted values, displayed in Figure 3.3.3.1 [(a)-(f)], there does not appear to be any evidence of trends or patterns present; hence, the final linear models fitted to the three river and three lochs sites seem to be appropriate.

However, when taking into account the adjusted R-squared values of each of the final models fitted, there appears to be a lot of unexplained variation. The final models are a fairly poor fit to the data at Callater Burn and the River Tweed above Gala Water Foot with adjusted R-squared values of 50.5% and 43%, respectively. This is similar in all 3 loch sites, with adjusted R squared values of 21.6%, 39.3% and 13% at the Lochs Naver, Kilbirnie –Beith and Lomond – Creinch, respectively. The only site, which the model seemed to fit adequately, was at the river site, Dall Bridge, which had an adjusted R squared value of 78.4%. Therefore, other models shall be explored, in an attempt to find a better model, which appropriately describes the behaviour of log TOC at each of the sites.

Summary of Final Linear Models Fitted to Sites					
	Sites	Terms	Estimate	St. Error	Pr(> t )
Rivers	Callater Burn	Intercept	-70.41	11.89	<0.001
		Year	0.04	0.01	<0.001
		Cos(Month)	-0.15	0.05	<0.001
		Sin(Month)	-0.38	0.04	<0.001
		Log(Flow)	0.45	0.05	<0.001
	Dall Bridge	Intercept	2.51	0.06	<0.001
		Cos(Month)	-0.29	0.06	<0.001
		Sin(Month)	-0.44	0.05	<0.001
		Log(Alkalinity)	-0.39	0.05	<0.001
	Tweed above Gala Waterfoot	Intercept	5.24	0.77	<0.001
		Cos(Month)	-0.16	0.05	0.002
		Sin(Month)	-0.23	0.04	<0.001
		Log(Alkalinity)	-0.86	0.22	<0.001
Lochs	Kilbirnie (Beith)	Intercept	-98.4	26.25	<0.001
		Year	0.05	0.01	<0.001
		Cos(Month)	0.04	0.06	<0.001
		Sin(Month)	-0.36	0.05	<0.001
		Log(Alkalinity)	-0.15	0.05	<0.001
	Lomond (Creinch)	Intercept	-42.27	13.83	0.003
		Year	0.02	0.006	0.008
	Loch Naver	Intercept	3.76	0.67	<0.001
		Cos(Month)	-0.06	0.05	0.29
		Sin(Month)	-0.12	0.05	0.02
		Log(Alkalinity)	-0.34	0.14	0.02

**Table 3.3.3.1: Summary of the final linear models fitted to each sites; and the significance of each term when included in the final linear models.**



**Figure 3.3.3.1: Residuals vs Fitted values plots for the final linear models fitted to Callater Burn(a), Loch Naver (b), Dall Bridge (c), Loch Kilbirnie (d), Tweed above Gala Waterfoot (e) and Loch Lomond (f).**

### 3.3.4 Additive Models and Non-Parametric Regression

Multiple linear regression models were fitted in the previous sub-section; however, after inspecting the various plots, it seems plausible that non-parametric regression techniques may be more appropriate. A generalized additive model (Hastie and Tibshirani, 1986 and 1990; Wood, 2006) [which is fitted using the back-fitting algorithm] may be more suitable, as it is more flexible than linear models. For example, the trend over time is not always linear - additive models fit a smooth curve to the data, which effectively captures the shape of the data over time, when the trend does not appear to be linear. Additive models are a non-parametric regression technique, and shall be explored in this section.

The linear regression models explored previously can be extended to additive models in the following manner:

$$y_i = \beta_0 + m_1(x_{1i}) + \dots + m_p(x_{pi}) + \varepsilon_i, \quad i=1, \dots, n. \quad (3.3.4.1)$$

The  $m_j$  are functions whose shapes are unrestricted, apart from an assumption of smoothness and the constraint, for identifiability, that  $\sum_{i=1}^n m_j(x_{ji}) = 0$  for all  $j=1, \dots, p$ . As a consequence, we usually estimate  $\beta_0$  by  $\bar{y}$ . This allows a very flexible set of modelling tools. To see how these models can be fitted, consider the case of only two covariates,

$$y_i = \beta_0 + m_1(x_{1i}) + m_2(x_{2i}) + \varepsilon_i, \quad i=1, \dots, n. \quad (3.3.4.2)$$

A rearrangement of this as  $y_i - \beta_0 - m_2(x_{2i}) = m_1(x_{1i}) + \varepsilon_i$  suggests that an estimate of component  $\hat{m}_1$  can then be obtained by smoothing the residuals of the data after fitting  $\hat{m}_2$ . If we express the curve estimator in symbolic form as  $\hat{m} = Sy$ .

Where  $\hat{m}$  denotes the vector of estimates at a set of evaluation points of interest,  $S$  denotes a smoothing matrix whose rows consist of the weights appropriate to estimation at each evaluation point, and  $y$  denotes the observed responses in vector form. Then,

$$\hat{m}_1 = S_1(y - \bar{y} - \hat{m}_2) \quad (3.3.4.3)$$

and similarly, subsequent estimates of  $\hat{m}_2$  can be obtained as

$$\hat{m}_2 = S_2(y - \bar{y} - \hat{m}_1). \quad (3.3.4.4)$$

These smoothing operations are repeated until convergence. In general, a model with  $p$  covariates, like so:

$$y_i = \beta_0 + m_1(x_{1i}) + \dots + m_p(x_{pi}) + \varepsilon_i, \quad i=1, \dots, n, \quad (3.3.4.5)$$

is a simple extension of the steps outlined for two covariates gives a form of the backfitting algorithm. At each step we smooth over a particular variable using as a response the  $y$  variable with the current estimates of the other components subtracted. The backfitting algorithm can be expressed as:

$$\hat{m}_j^{(r+1)} = S_j \left( y - \hat{\beta}_0 - \sum_{k < j} \hat{m}_k^{(r+1)} - \sum_{k > j} \hat{m}_k^{(r)} \right) \quad j=1, \dots, p \quad (3.3.4.6)$$

(Nobile and Bowman, 2010; Wood, 2006; Hastie and Tibshirani, 1986 and 1990)

### 3.3.5 Fitting Additive Models to Sites

Similar to section 3.3.1, additive models were fitted at each of the sites, initially to consider trend and seasonality only. The term year and the covariates shall be expressed in the additive model in the same manner as before, but, the harmonic terms (sine and cosine) shall not be included – the seasonality shall be represented by the month (i.e. 1,2,...,12) in which the sample was taken. At first, the following additive model was fitted to each site:

$$y_i = \beta_0 + m_1(\text{year}_i) + m_2(\text{month}_i) + \varepsilon_i, \quad (3.3.5.1)$$

Where  $y_i$  is the level of log TOC and  $\varepsilon_i$  are assumed to be independent with mean 0 and constant variance. The term, month, in the model is fitted using a cyclic cubic regression spline as a base to ensure that the start point is the same as the end point (also known as a ‘circular’ term). Furthermore, the degree of smoothing applied to each smooth term in the model is chosen by a method known as Generalized Cross Validation (GCV). The additive model (3.3.5.1) fitted to Loch Kilbirnie is displayed in Figure 3.3.5.1 [(a) and (b)]. The plot highlights that the model effectively captures the shape of the trend of log TOC at Loch Kilbirnie, but also the seasonal pattern. Both terms being highly significant in the model, with p-values less than 0.001. The additive model fitted to Loch Kilbirnie, has an adjusted R-squared value of 49.3% - this model already explains more of the variation in the data, than the final linear model fitted previously (as seen in Table 3.3.3.1).

Similar to before, covariates shall be included in the additive model to try to improve the trend and seasonality model already fitted. Additive models are flexible in the way that they allow covariates to be included as either a smooth term or as a linear term. For example, the covariate temperature could be included in the model (3.3.5.1) resulting in either an additive or additive semi-parametric model, like so:

Additive model: 
$$y_i = \beta_0 + m_1(\text{year}_i) + m_2(\text{month}_i) + m_3(T_i) + \varepsilon_i.$$

Additive Semi-parametric model: 
$$y_i = \beta_0 + m_1(\text{year}_i) + m_2(\text{month}_i) + \beta_1 T_i + \varepsilon_i,$$



To find the best additive model, firstly, at each of the river and loch sites, an additive model was fitted, which included the terms year and month, as well as the covariates temperature, pH, log(alkalinity), log(sulphate), log(nitrate) and log(flow) [if flow data was available]. The additive models (including all terms) were expressed as:

$$y_i = \beta_0 + m_1(\text{year}_i) + m_2(\text{month}_i) + m_3(A_i) + m_4(T_i) + m_5(\text{pH}_i) + m_6 \log(N_i) + m_7 \log(S_i) + m_8 \log(\text{Flow}_i) + \varepsilon_i. \quad (3.3.5.2)$$

Terms that were not significant (i.e. p-value not less than 0.05) were removed from the additive model. Figure 3.3.5.1 [(c)-(e)] displays the effect plots of the significant terms included in the following additive model fitted to the River Tweed above Gala Water Foot:

$$y_i = \beta_0 + m_1(\text{month}_i) + m_2(A_i) + m_3(S_i) + \varepsilon_i. \quad (3.3.5.3)$$

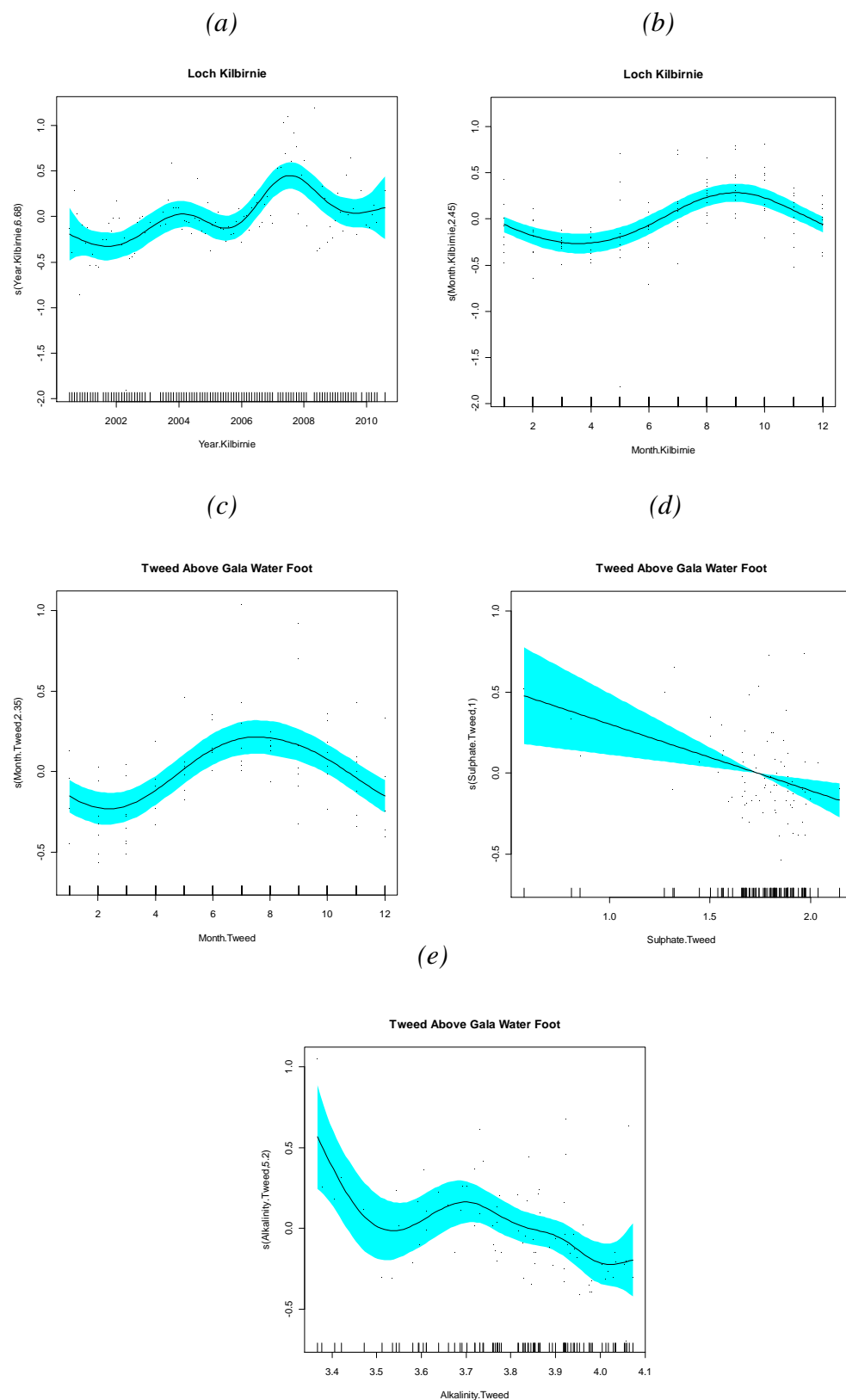
The plots in Figure 3.3.5.1 include the partial residuals from the GAM model fitted, as well as  $\pm 2$  standard error bands. However, it is clear from the effect plots displayed, as log Sulphate levels increase at the River Tweed, the levels of log TOC decreases in a linear manner. A favourable attribute of the *mgcv* package allows the additive model to be re-fitted as an additive semi-parametric model (including log Sulphate as a linear term):

$$y_i = \beta_0 + m_1(\text{month}_i) + m_2(A_i) + \beta_1 S_i + \varepsilon_i. \quad (3.3.5.4)$$

Whether a term should be included in the GAM model as a parametric or non-parametric term, an approximate F-test (Hastie and Tibshirani, 1990) can be used to formally test what would be more appropriate. The approximate F-test rejects expression (3.3.5.3) in favour of expression (3.3.5.4) [a p-value of <0.001 rejecting the ‘smooth’ term in favour of the parametric term].

The final additive models fitted to the river and loch sites are summarized in Tables 3.3.5.1 to 3.3.5.6. The final additive models highlight that the covariates only seem to have an affect on log TOC levels at the river sites. The additive models seem to be a good fit to the data at the river sites Callater Burn and Dall Bridge: adjusted R squared values of 73.6%, and 80%, respectively; and Figure 3.3.5.2 [(a) and (c)] displays that the residuals vs fitted values plots

show no trend or pattern. The additive model fitted to the River Tweed above Gala Waterfoot explains more of the variation than the linear model fitted previously, with an increased adjusted R squared value of 49.6%. The only significant terms included in the final additive models at the loch sites are either year, month or both. The three additive models fitted to the loch sites seem to be a poor fit to the data, with adjusted R-squared values 23.7% (Loch Naver), 49.3% (Loch Kilbirnie –Beith) and 10.7% (Loch Lomond – Creinch); although, Figure 3.3.5.2 [(b), (d) and (e)] reveals reasonable residuals vs fitted values plots. This suggests that there are other covariates possibly influencing the levels of log TOC at these particular loch sites. These models re-iterate the point, that it seems plausible, that across Scotland, the log TOC levels in rivers are affected by different physical and chemical factors than lochs. But, only six sites have been investigated in great detail in this chapter. They are not a reflection of all of the Scottish rivers and lochs – just an insight.



**Figure 3.3.5.1: Effect plots of additive model fitted at: the River Tweed above Gala Waterfoot [(a)- (c)]; and Loch Kilbirnie [(d) and (e)].**

Summary of Additive Semi-Parametric Model Fitted at Callater Burn			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	0.89	0.07	<0.001
Temperature	0.02	0.009	0.0192
Smooth Terms	Npar Df	Npar F	Pr(F)
Year	5.21	14.58	<0.001
Month	2.79	17.96	<0.001
Log(Flow)	3.01	31.88	<0.001
Log(Sulphate)	6.97	2.1	0.04

**Table 3.3.5.1: The significance of each term, when included in the final additive semi-parametric model, at the River site Callater Burn. Note: - ‘Npar Df’ refers to non-parametric degrees of freedom; ‘Npar F’ refers to non-parametric F-value.**

Summary of Additive Model Fitted at Dall Bridge			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	2.23	0.03	<0.001
Smooth Terms	Npar DF	Npar F	Pr(F)
Month	3.56	18.67	<0.001
Log(Alkalinity)	2.02	29.78	<0.001

**Table 3.3.5.2: The significance of each term when included in the final additive model at the River site Dall Bridge.**

Summary of Additive Semi-parametric Model Fitted at Tweed above Gala Water Foot			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	2.03	0.23	<0.001
Log(Sulphate)	-0.41	0.13	0.002
Smooth Terms	Npar DF	Npar F	Pr(F)
Month	2.35	7.96	<0.001
Log (Alkalinity)	5.19	3.63	0.003

**Table 3.3.5.3: The significance of each term when included in the final additive semi-parametric model at the River site Tweed above Gala Waterfoot.**

Summary of Additive Model Fitted at Loch Naver			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	2.09	0.03	<0.001
Smooth Terms	-0.06	0.05	0.29
Month	-0.12	0.05	0.02

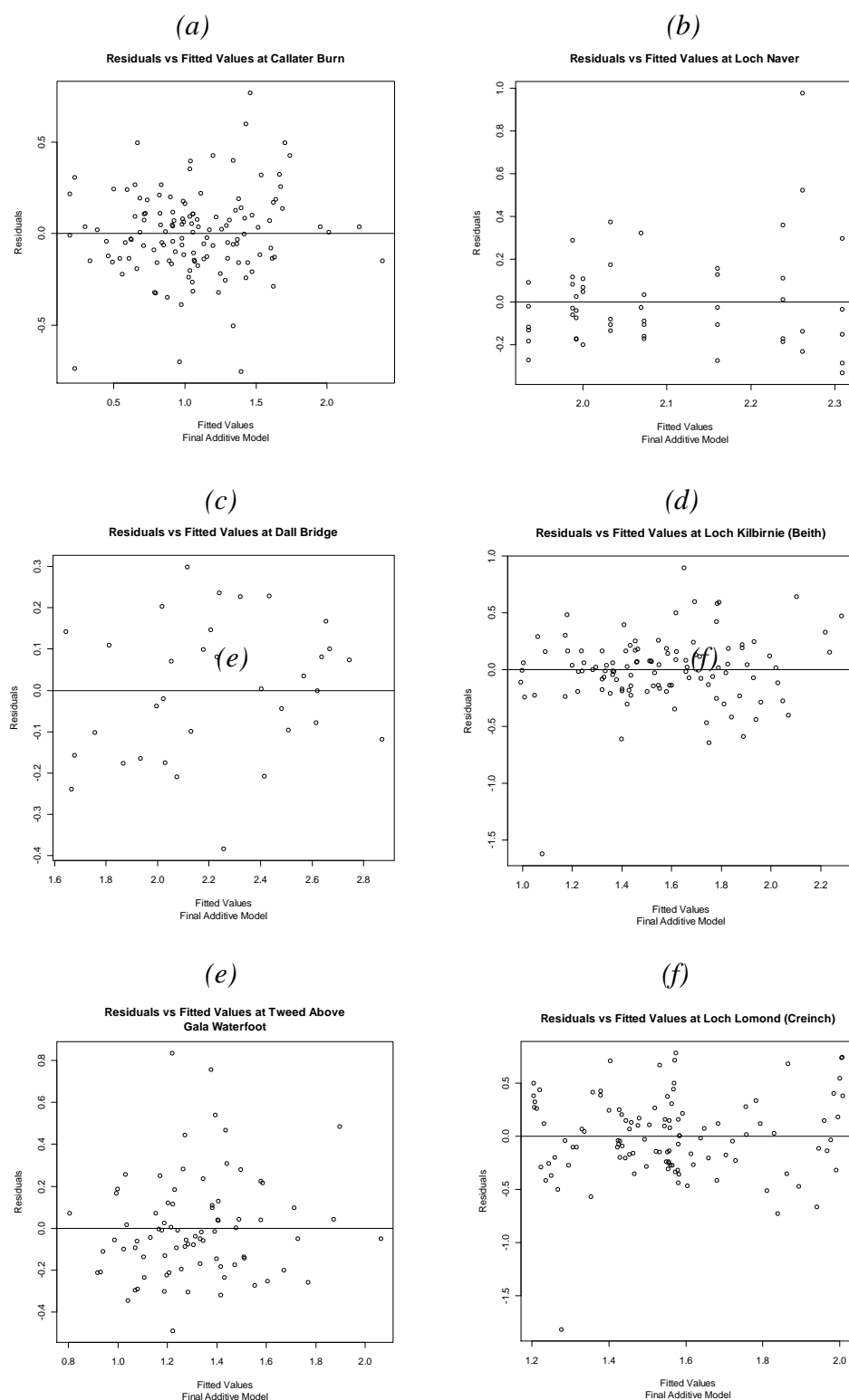
**Table 3.3.5.4: The significance of each term when included in the final additive model fitted at the site.**

Summary of Additive Model Fitted at Loch Kilbirnie (Beith)			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	1.56	0.03	<0.001
Smooth Terms	Npar Df	Npar F	Pr(F)
Year	6.68	7.79	<0.001
Month	2.45	3.25	<0.001

**Table 3.3.5.5: The significance of each term when included in the final additive mode fitted at the site Loch Kilbirnie (Beith).**

Summary of Additive Model Fitted at Loch Lomond (Creinch)			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	1.46	0.04	<0.001
Smooth Terms	Npar Df	Npar F	Pr(F)
Year	1.43	5.14	0.01

**Table 3.3.5.6: The significance of each term when included in the final additive model fitted at the site Loch Lomond (Creinch).**



**Figure 3.3.5.2: Residuals vs Fitted values plots for the final additive models fitted to Callater Burn(a), Loch Naver (b), Dall Bridge (c), Loch Kilbirnie (d), Tweed above Gala Waterfoot (e) and Loch Lomond (f).**

### 3.4 Choosing The ‘Best’ Model: Linear or Additive?

To formally test whether one model is better than another, additive models can be compared to linear models using F-tests (Hastie and Tibshirani, 1990; Bowman and Azzalini, 1997). Hastie and Tibshirani (1990) recommend the use of residual sums-of-squares and their associated degrees of freedom to provide guidance for model comparisons. For an additive model, the residual sum-of-squares can easily be defined as

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (3.3.6.1)$$

Where  $\hat{y}_i$  denotes the fitted value, produced by evaluating the additive model at the observation  $x_i$ . Comparisons of a linear model to an additive model can be expressed quantitatively as

$$F = \frac{(RSS_2 - RSS_1)/(df_2 - df_1)}{RSS_1 / df_1}, \quad (3.3.6.2)$$

by analogy with the F-statistic used to compare linear models. Unfortunately, this analogy does not extend to distributional calculations. However, Hastie and Tibshirani (1990) suggest that at least some approximate guidance can be given by referring the observed nonparametric F-statistic to an F-distribution with  $(df_2 - df_1)$  and  $df_1$  degrees of freedom.

The null hypothesis is that the linear model (the simpler model) is an adequate fit to the log TOC levels at that particular site; the alternative hypothesis being that the additive model is a better fit to the data. If the F-statistic is greater than the rejection region found using the F-distribution, the null hypothesis is rejected in favour of the alternative. A summary of the approximate F-tests for each of the river and loch sites can be seen in Table 3.3.6.1.

With regards to the river sites, it is apparent from Table 3.3.6.1 that the sites (Callater Burn and River Tweed) with a longer time series, are more appropriately described by an additive model. This was expected, as the trends displayed by these sites, did not behave in a linear manner. However, the site Dall Bridge could be described adequately using a linear model.



On the other hand, the loch sites did not seem to show as clear a pattern – Loch Lomond (Creinch) with the longest time series, was better described by a linear model, than an additive model. The lochs seem to be site specific with regards to way in which the levels of log TOC behaves, and the way in which they are most effectively modelled.

Comparison of Final Linear and Additive Models Fitted to River and Loch Sites Using Approximate F-tests						
Site	Model	RSS	Df	F- statistic	Rejection Region	Preferred Model
R. Callater Burn	Linear	10.59	124	4.47	1.81	Additive
	Additive	7.11	111.03			
R. Dall Bridge	Linear	1.03	30	1.87	3.09	Linear
	Additive	0.88	27.43			
R. Tweed	Linear	5.51	76	3.19	2.41	Additive
	Additive	4.58	71.45			
L. Kilbirnie	Linear	12.69	87	1.92	1.68	Additive
	Additive	3.64	38.02			
L. Creinch	Linear	8.85	83	2.39	3.38	Linear
	Additive	9.26	84.57			
L. Naver	Linear	2.84	47	2.38	4.18	Linear
	Additive	2.71	46.09			

**Table 3.3.6.1: Comparison of final linear and additive models fitted to the river and loch sites using an Approximate F-test.**

### 3.5 Conclusion

The main aim of this chapter was to explore three river and three loch sites in detail. The sites were chosen on the basis that each site represented the different lengths of time series present in the whole data set. The trend, seasonality and relationships with covariates at each site were examined. The overall aim of this chapter was explore suitable methods of modelling log TOC at individual sites.

At Callater Burn, (the river site with the longest time series), the log TOC levels appear to increase from the early 1980's until the early 2000's – after 2004, the log TOC levels seem to “level off”. However, the log TOC levels in the other river sites, Tweed above Gala Water Foot and Dall Bridge, (with shorter time series) remain fairly flat across the years. The loch sites show a similar trend to the three river sites: log TOC seems to increase from the early 1990's up until the mid-2000's. However, Loch Naver with only five years of data, shows no significant trend, and behaves similar to the other two loch sites in their latter years.

With regards to seasonality, log TOC seems to follow a seasonal pattern in all three river sites and Loch Kilbirnie. At these sites, it seems that levels of log TOC appear to increase from early spring up until early autumn – during late autumn and winter, the log TOC levels seem to decrease. There does not seem to be a strong seasonal pattern in either Loch Lomond or Loch Naver.

From the exploratory plots, an initial impression of the relationships between log TOC and the different covariates could be formed. At the river sites Tweed and Dall Bridge, an increase in the log Alkalinity levels was associated with a decrease in log TOC levels. An increase in temperature was associated with an increase in log TOC levels at each of the three river sites. Log nitrate seemed to be associated with a decrease in log TOC levels at the River Tweed only. An increase in log flow seemed to be associated with an increase in log TOC levels at Callater Burn. On the other hand, the covariates did not appear to have a strong relationship with log TOC at any of the loch sites. If anything, an increase in

temperature and log alkalinity seemed to be associated with an increase in log TOC – but, this was a very weak relationship

Based on the six sites investigated, the exploratory analysis suggested that the covariates were more likely to be associated with a change in log TOC levels in rivers, than lochs. But, it was important to remember that only three river and three loch sites were being considered.

Having explored the trend, seasonality and relationship with covariates, different modelling techniques were applied to each site, separately. Linear models and generalized additive models were explored – each model addressing trend, seasonality and the covariates. A linear model and generalized additive model was fitted to each site.

From the linear regression, the rate of increase in log TOC at sites could be calculated (note: it was only calculated for sites with a significant trend term). The levels of log TOC at Callater Burn for any given month, on average, are increasing by 0.04 mg/l for every one year increase at Callater Burn; and for any given month, on average, the level of log TOC is increasing by 0.02 mg/l at Loch Lomond (Creinch) and increasing by 0.06 mg/l at Loch Kilbirnie (Beith), for every one year increase. As discussed in Chapter 1, according to Moxley (2010), the rate of TOC increase, averaged across all sites with increasing concentrations, was 0.12 milligrams per litre per year (mg/l/y). Hence, the rate of increase does not seem to be as severe at these select sites.

F-tests (Hastie and Tibshirani, 1990) were used to formally compare the inclusion of a term as being linear or non-parametric in a GAM; but also, an F-test (Hastie and Tibshirani, 1990; Bowman and Azzalini, 1997) was used to compare whether a linear or additive model was more appropriate for describing the behaviour of log TOC at each of the sites.

The length of time period did not seem to determine whether a linear or additive model was a more appropriate fit to a site. The river sites Callater Burn and River Tweed (with longer time series than the other site, Dall Bridge) were appropriately described by an additive model. This was expected, as the trends displayed by these sites, did not behave in a linear

manner. However, Loch Lomond (Creinch) with the longest time series (out of the three lochs), was more appropriately described by a linear model. Based on these six sites, it seems that the most appropriate modeling technique is specific to each site.

This chapter has considered three river and three loch sites –these sites are not spatially grouped or ecologically connected; but, have provided a further insight into the trend, seasonality and relationships of log TOC. The next chapter shall consider sites which are located in the same river network. The relationship between their spatial location, distance between sites, and the way in which the river flow connects each site, shall have to be considered in order to find a suitable model to capture the behaviour of log TOC.

# Chapter 4

## River Networks

SEPA has implemented the River Basin Management Plan (2009-2015) in accordance with the Water Framework Directive (2000) to ensure that Scotland maintains or takes steps to move towards good water quality in all water bodies. For monitoring purposes, Scotland is split up into different catchments. Catchments include all of the rivers, lochs, wetlands and groundwater which eventually drain into the sea, as well as coastal waters and estuaries. River catchments are made up of different tributaries. The term ‘river network’ is used to describe a particular region within a catchment. It is thought that, all the different tributaries included within a given network can influence the levels of total organic carbon across the whole river network. This is why river networks are investigated as a whole. We shall focus on the River Dee, situated in Aberdeenshire, which rises in the Cairngorms and flows North-East across Scotland towards the North Sea as displayed in Figure 4.1. At first, the focus shall be on the sites which sit on the main river channel (i.e. those sites which are located on the River Dee itself) then the focus shall be switched to the River Dee network as a whole (which includes rivers and streams which flow into the River Dee). The main aims of this chapter are: to model those sites which are situated on the River Dee’s main channel individually; find a global model which best describes all of the sites situated on the River Dee’s main channel; predict the levels of log TOC across the whole river network based on

the information available; and to find a model which captures the behaviour of log TOC across space and time in the river network.

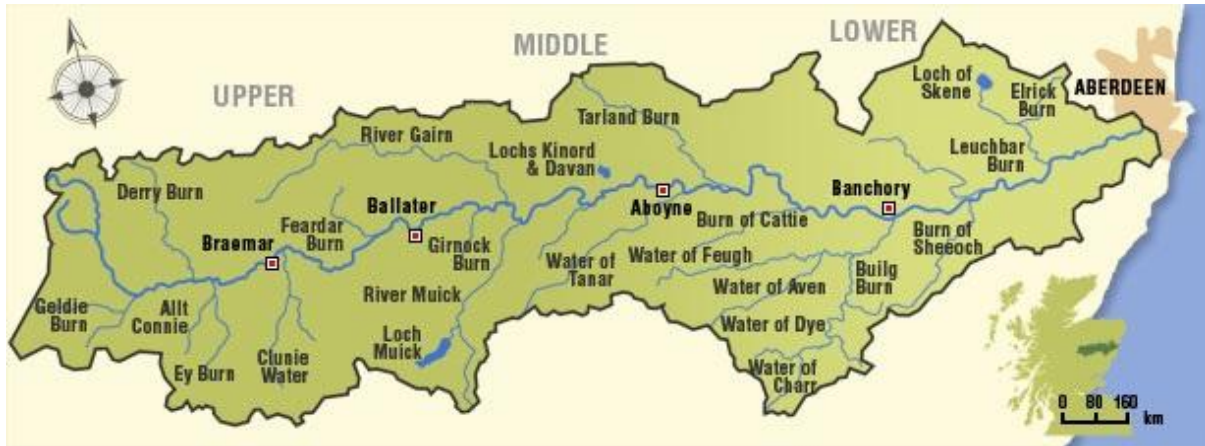
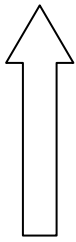


Figure 4.1: Location of the River Dee (River Dee Map - <http://www.theriverdee.org/explore-the-catchment.asp>)

## 4.1 Initial Impression of the Sites Along the Main Channel (i.e. The River Dee)

Similar to the previous sections, to gain an initial impression of each of the five sites situated on the main river, the log TOC at each site shall be plotted against Year, Month, log Alkalinity, pH, log river Flow, Temperature, log Nitrate and log Sulphate. The sites being explored in sections 4.1, 4.2 and 4.3 are summarised in Table 4.1.1 and the corresponding time series are plotted in Figure 4.1.1 (a). The time series plot highlights the missing data of sites 1, 2, 4 and 5 (particularly) between the year 2000 and 2007. Figure 4.1.1 features the seasonality of each site (b) and also scatterplots of log TOC against different covariates [(c)-(e)]. Sites 1, 2, 4 and 5 (between the years 1990 and 2000) seem to follow the same trend –

all sites revealing an increase in levels of log TOC up until 2000. Site 5 with the longest times series shows the “leveling off” of log TOC levels post 2000. The time series plot highlights the small amount of data available for the site ‘Banchory Bridge’ and the large amount of missing data between 1990 and 2007. The colour scheme in the time series plot effectively highlights that as the water flows towards the North Sea i.e. down the river, the levels of log TOC seem to increase.

Flow Direction	Site	Site Name	Time Period	Covariate Data Not Available
	1	Bridge of Dee	1989-2000	Nitrate
	2	Milltimber	1989-2010 ( large gap with missing data)	
	3	Banchory Bridge	1989–2010 (large gap between 1991 and 2007)	Flow,Nitrate, Sulphate
	4	Potarch Bridge	1989-2010	Nitrate
	5	Linn of Dee	1989-2000	

**Table 4.1.1: Summary of the 5 river sites under investigation situated on the River Dee**

The seasonality plots in Figure 4.1.1 (b) suggest that in all 5 sites, the seasonal pattern is very similar to those sites investigated in previous sections– there is an increase in levels of log TOC during the summer and autumn months, followed by a decrease in the winter.

Considering the log alkalinity levels displayed in Figure 4.1.1 (c), it appears that site 5 seems to have the lowest levels of log alkalinity. It seems plausible that the level of alkalinity in the water increases as the water moves downstream. But, as the alkalinity increases at site 5, the log TOC also increases. For sites 1,2 and 4, it appears that an increase in the level of log alkalinity, is associated with a decrease in the level of log TOC.

The plot of log TOC against river flow in Figure 4.1.1 (d) highlights, at sites 1,2,4 and 5, that an increase in river flow induces an increase in log TOC levels. Site 5 again, stands out from the other sites, as it seems to have a lower volume flow than the others [Linn of Dee is a long way upstream of the other sites and is very narrow (but quite deep) so has a lower

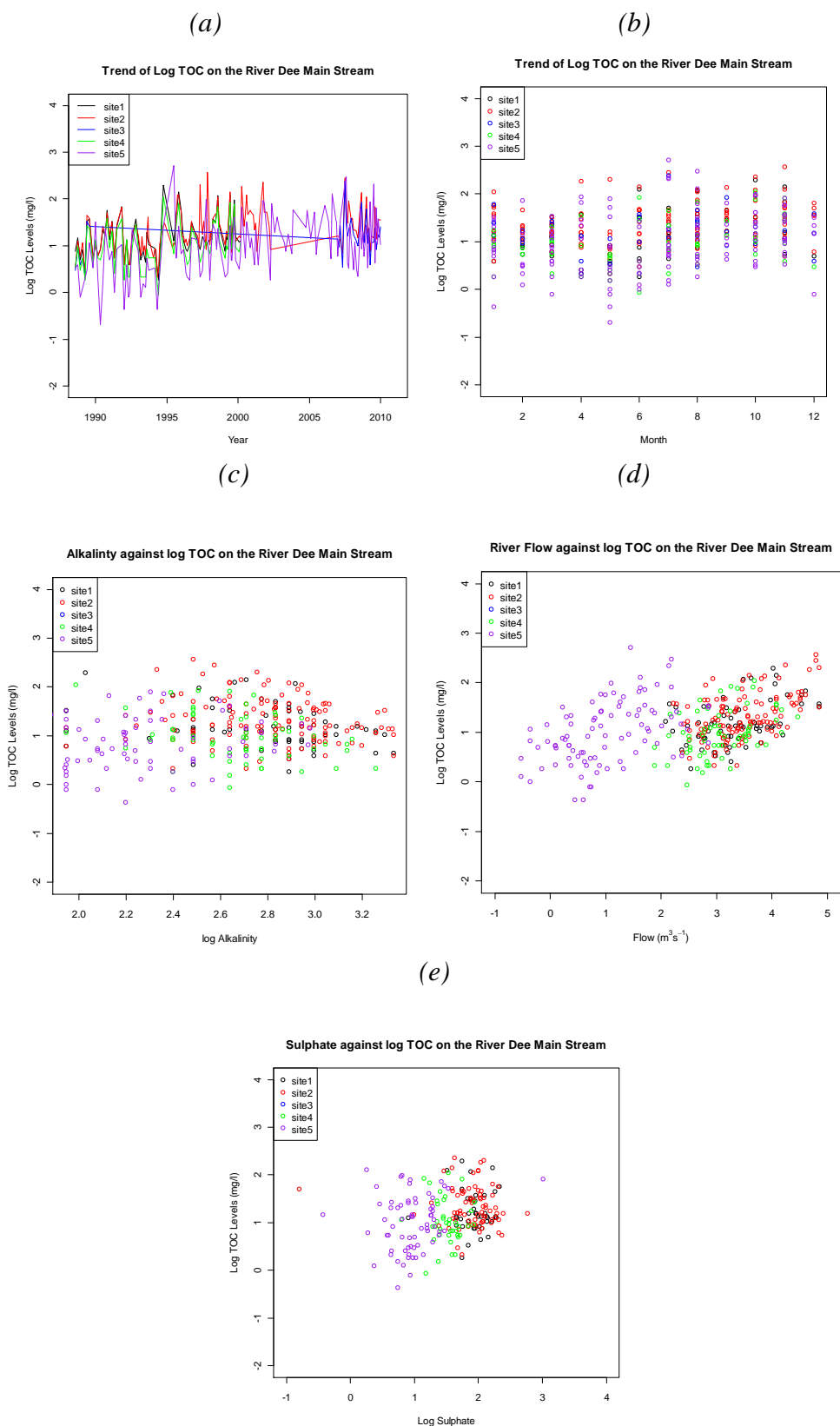
volume than other sites], which in turn possibly explains the lower levels of log TOC present at site 5.

Considering Figure 4.1.1 (*e*), it is evident that levels of log sulphate seem to be lowest, again, at site 5. Looking at sites 1, 2, 4 and 5, it seems likely that an increase in the level of log sulphate is associated with an increase in log TOC levels.

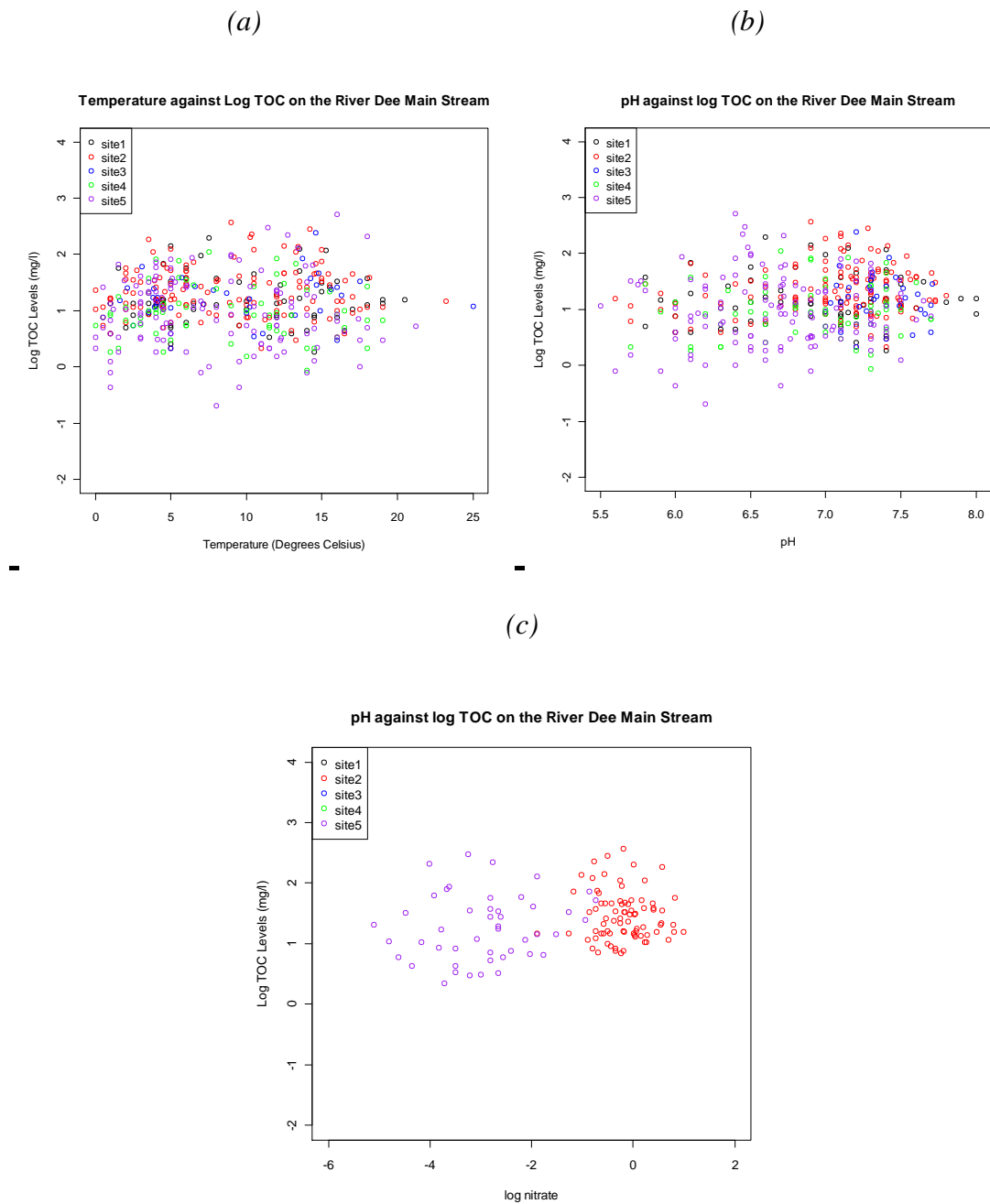
With regards to the temperature, pH and log nitrate levels at each of the sites, these covariates did not appear to have any significant effect on the 5 River Dee sites as Figure 4.1.2 [(a)-(c), respectively] displays.

For the purpose of Sections 4.2 and 4.3, the decision was made not to include Banchory Bridge in the analysis as it does not seem like a worthwhile exercise, based on the missing total organic carbon data, and the small amount of data (if any) available for the covariates.





**Figure 4.1.1: Log TOC plotted against Year (a), Month (b), Log Alkalinity (c), Log Flow (d) and Log Sulphahte (e) at the 5 river sites situated on the River Dee**



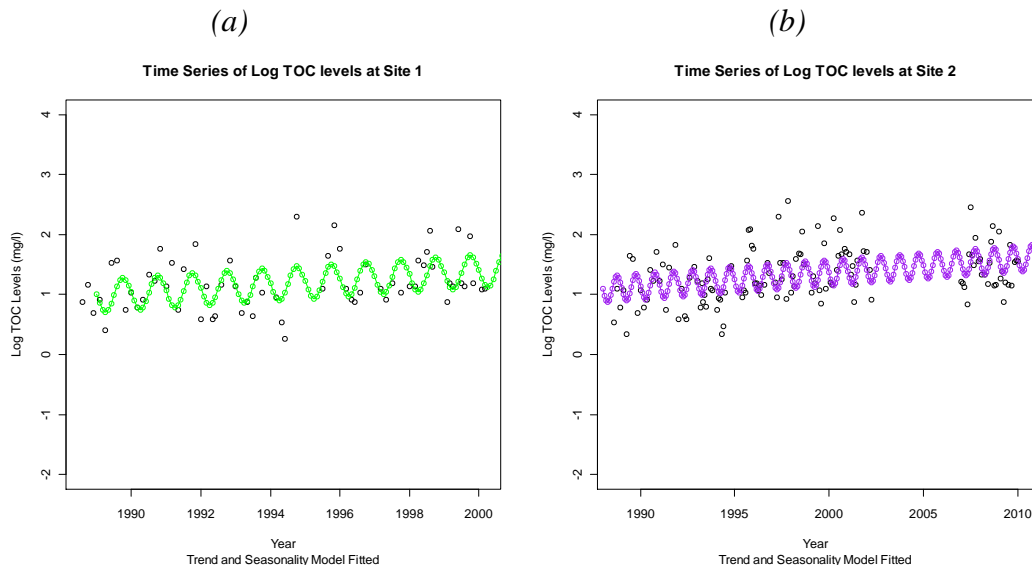
**Figure 4.1.2: Log TOC plotted against Temperature (a), pH (b) and log nitrate (c) at the 5 river sites situated on the River Dee.**

## 4.2 Modelling Each Site Along the Main Channel (i.e. The River Dee)

Firstly, each of the sites shall be modelled independently of one another, to gain an insight into behaviour of log TOC at each site, before attempting to model the sites on the River Dee as a whole. As mentioned previously, the site “Banchory Bridge” shall not be explored in this section. This section shall focus on modelling the following sites only: Bridge of Dee, Milltimber, Potarch Bridge and Linn of Dee. Similar to the previous chapter, linear models and generalized additive models shall be fitted to each of the sites to capture the behaviour of log TOC. Again, an F-test (Hastie and Tibshirani, 1990; Bowman and Azzalini, 1997) shall be used to compare the linear and additive models fitted to each site.

Similar to Section 3.3.1, linear models were fitted to each of the sites. At first, a linear model considering the trend and seasonality (using harmonic terms) only was fitted to each of the sites using equation (3.3.1.5); this linear model was then extended to a multiple linear regression which included all of the available covariates, as linear terms, using equation (3.3.3.1). Figure 4.2.1 displays the trend and seasonality models fitted to the time series at sites 1 and 2 – it can be seen visually, that there is a lot of unexplained variation, which is reinforced by the adjusted R-squared values of 26.3% and 22.8% (respectively).

This was also the case at sites 4 and 5, with adjusted R-squared values of 21.4% and 25.9%, respectively. At each of the sites, the linear models (which initially included trend and seasonality terms) were improved by including one or more of the covariates. The final linear models fitted to sites 1, 2, 4 and 5 are summarised in Table 4.2.1. Note: the correlation of the residuals was explored, again, using Auto Correlation Function Plots similar to those seen in Chapter 3 [Figure 3.3.2.1 for example] - there did not appear to be any significant correlation between the residuals; hence, correlation did not need to be incorporated in the final linear models.



**Figure 4.2.1: Time series of Log TOC at sites 1 (a) and 2 (b) on the River Dee, with the corresponding trend and seasonality model fitted to each plot.**

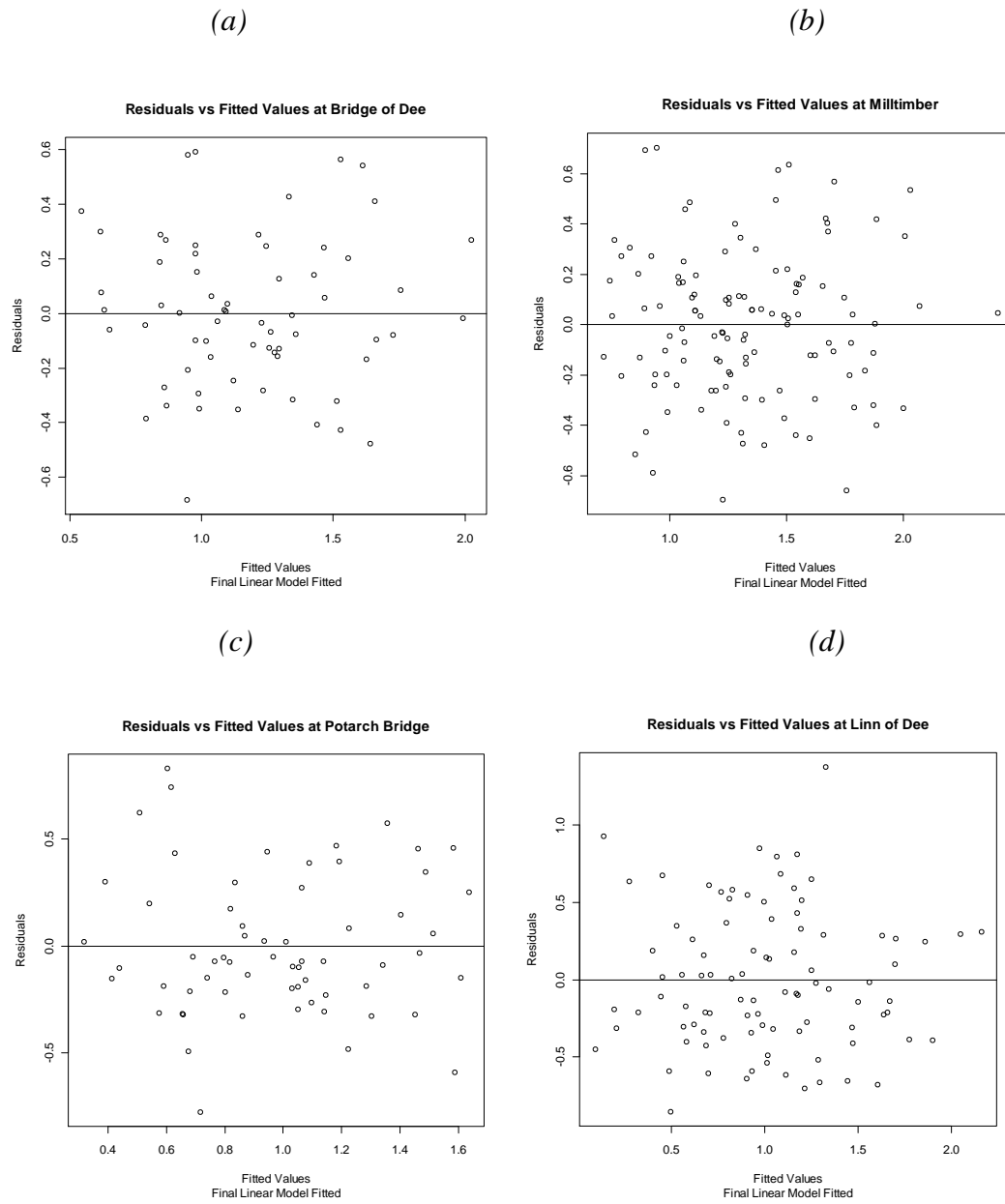
As expected, the trend and seasonality are significant in the final linear models fitted at each of the 4 River Dee Sites. However, it is also of interest to note that the covariate ‘log flow’ has a significant effect on log TOC levels at each of the 4 sites (with the addition of log Alkalinity at Bridge of Dee).

The residuals vs fitted values from the final linear models are displayed in Figure 4.2.2 [(a)-(d)] – there does not appear to be any strong trend or pattern; hence, the linear models seem to be a reasonable fit to the data. However, the final linear models fitted to the sites Bridge of Dee, Milltimber, Linn of Dee and Potarch Bridge had the following adjusted R-squared values: 56.4%, 56.2%, 47.3% and 48.6%, respectively. The adjusted R-squared values suggest that the final linear models fitted have left a lot of unexplained variation. It is therefore of interest to investigate if a different modelling approach would be more appropriate. Non-parametric regression may be more appropriate for capturing the behaviour of the trend, seasonality and covariates at each of the sites. Hence, using methods explored in section 3.3.5, additive models were also fitted to each of the 4 sites on the River Dee in an attempt to improve the modelling at each site. The final additive models are summarised in Tables 4.2.2 to 4.2.5. Interestingly, sites 1 and 2 (Bridge of Dee and Milltimber) and sites 4

and 5 (Potarch Bridge and Linn of Dee) have the same covariates included in their final additive models [i.e. Year, Month, log Alkalinity, log Flow; and Year, Month, log Flow respectively].

<b>Summary of the Multiple Linear Regression</b> <b>Models Fitted to the River Dee's Sites on the Main Channel</b>							
<b>Site</b>	<b>Variables</b>						
		<b>Intercept</b>	<b>Year</b>	<b>Cosine</b>	<b>Sine</b>	<b>Log (Alkalinity)</b>	<b>Log (Flow)</b>
<b>Bridge of Dee</b>	<b>Coeff</b>	-69.39	0.04	-0.05	-0.37	-0.45	0.28
	<b>P-value</b>	0.001	0.001	0.44	<0.001	<0.001	<0.001
<b>Milltimber</b>	<b>Coeff</b>	-41.84	0.02	-0.99	-0.33	-	0.47
	<b>P-value</b>	<0.001	<0.001	0.03	<0.001	-	<0.001
<b>Potarch Bridge</b>	<b>Coeff</b>	-86.11	0.04	-0.067	-0.35	-	0.53
	<b>P-value</b>	<0.001	<0.001	0.38	<0.001	-	<0.001
<b>Linn of Dee</b>	<b>Coeff</b>	-64.57	0.03	-0.148	-0.386	-	0.476
	<b>P-value</b>	0.001	0.001	0.065	<0.001	-	<0.001

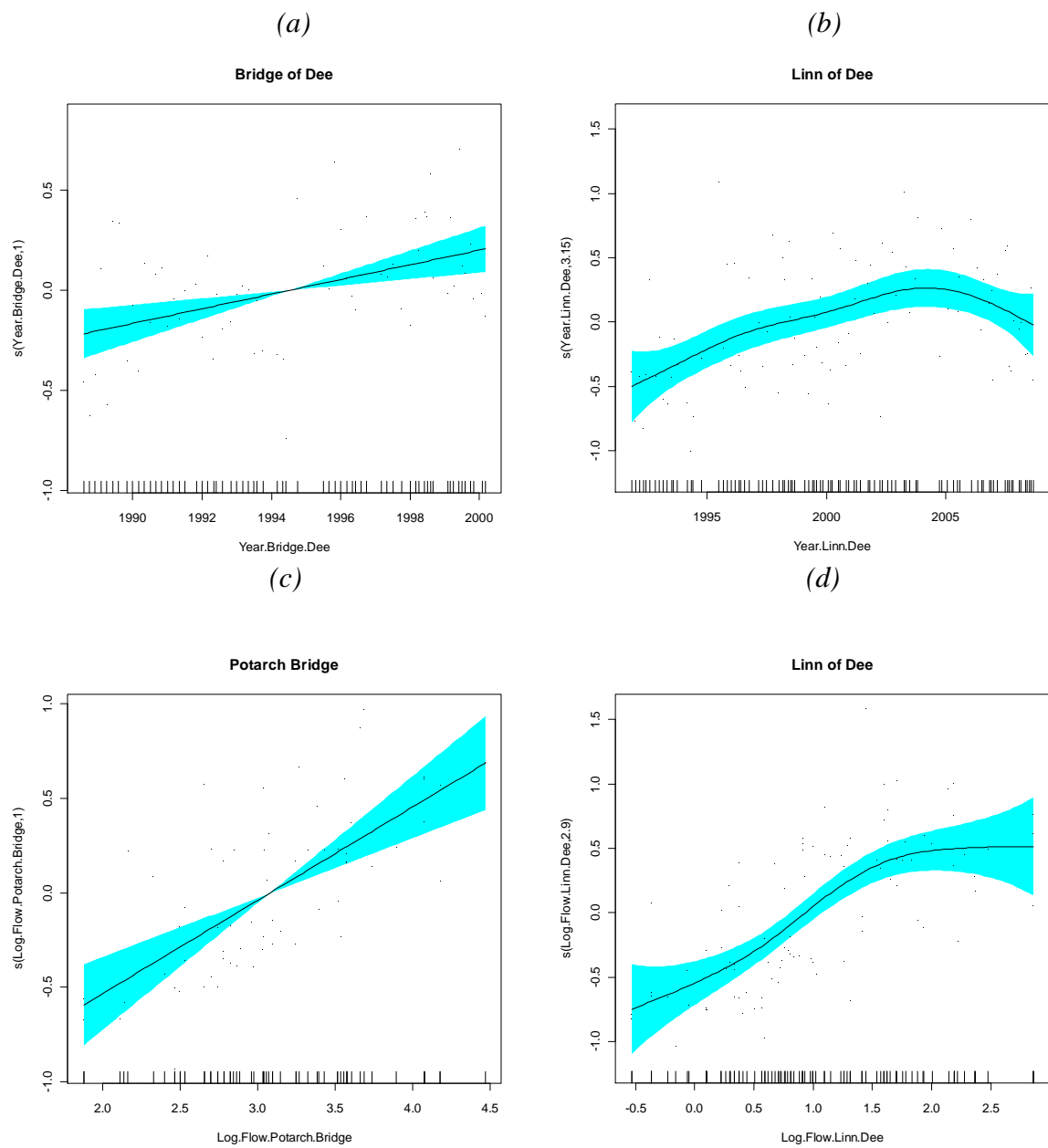
**Table 4.2.1: The significance of each term when included in the final linear models fitted to each of the sites.**



**Figure 4.2.2: Residuals vs Fitted Values plotted for the final linear models fitted to the sites Bridge of Dee (a), Milltimber (b), Potarch Bridge (c) and Linn of Dee (d).**

Figure 4.2.3 highlights the key difference between the final additive models fitted at Bridge of Dee (a) and Linn of Dee (b): the term ‘Year’ is fitted as a linear term at Bridge of Dee (results of the approximate F-test revealed a p-value of  $<0.001$ , hence rejecting an additive model including ‘Year’ as a smooth term); but, on the contrary, ‘Year’ was fitted as a smooth term at Linn of Dee - the approximate F-test rejecting the additive model including ‘Year’ as a linear term (p-value  $<0.05$ ) in favour of the additive model including ‘Year’ as a smooth term. A reason for this may be the differing time periods at each site – as seen in the exploratory analysis, the levels of log TOC increase in a linear fashion from early 1990’s to early 2000’s and then start to ‘level-off’. Both sites increase linearly between 1990 and 2000; but, the site ‘Linn of Dee’ with data post 2000 shows a decrease in log TOC levels after the year 2005. Hence, a smooth year term seems to be more appropriate for the longer time series, such as the Linn of Dee.

However, this explanation is not plausible for the difference in final additive models at Potarch Bridge and Linn of Dee: the term ‘log flow’ is included as a linear term at Potarch Bridge, but as a smooth term at Linn of Dee. An increase in the log flow levels reveal a slightly different effect on the log TOC at each site. Potarch Bridge showing a linear increase in log TOC; Linn of Dee showing an increase resembling an ‘S’ shape – this can be seen graphically in Figure 4.2.3 [(c) and (d)].



**Figure 4.2.3: A selection of effect plots from the final additive models fitted to sites Bridge of Dee (a), Linn of Dee (b), Potarch Bridge (c) and Linn of Dee (d).**



Summary of Additive Semi-parametric Model Fitted at Bridge of Dee			
Parametric Coefficients	<u>Estimate</u>	<u>Std. Error</u>	<u>Pr(&gt; t )</u>
Intercept	-71.8	20.41	<0.001
Year	0.04	0.01	<0.001
Smooth Terms	<u>Npar DF</u>	<u>Npar F</u>	<u>Pr(F)</u>
Month	2.47	13.8	<0.001
Log (Alkalinity)	1.83	4.36	0.01
Log ( Flow)	1.61	7.61	0.001

**Table 4.2.2: The significance of each term when included in the final additive semi-parametric model at the River site Bridge of Dee.**

Summary of Additive Semi-parametric Model Fitted at Milltimber			
Parametric Coefficients	<u>Estimate</u>	<u>Std. Error</u>	<u>Pr(&gt; t )</u>
Intercept	1.33	0.02	<2e-16
Year	0.02	0.005	<0.001
Smooth Terms	<u>Npar DF</u>	<u>Npar F</u>	<u>Pr(F)</u>
Month	3.19	17.6	<0.001
Log (Alkalinity)	2.29	3.12	0.03
Log ( Flow)	3.16	16.9	<0.001

**Table 4.2.3: The significance of each term when included in the final additive semi-parametric model at the River site Milltimber.**

Summary of Additive Semi-parametric Model Fitted at Potarch Bridge			
Parametric Coefficients	<u>Estimate</u>	<u>Std. Error</u>	<u>Pr(&gt; t )</u>
Intercept	-84.83	24.06	<0.001
Year	0.04	0.01	<0.001
Log (Flow)	0.49	0.09	<0.001
Smooth Terms	<u>Npar DF</u>	<u>Npar F</u>	<u>Pr(F)</u>
Month	2.12	7.78	<0.001

**Table 4.2.4:**The significance of each term when included in the final additive semi-parametric model at the River site Potarch Bridge.

Summary of Additive Model Fitted at Linn of Dee			
Parametric Coefficients	<u>Estimate</u>	<u>Std. Error</u>	<u>Pr(&gt; t )</u>
Intercept	1.01	0.043	<2e-16
Smooth Terms	<u>Npar DF</u>	<u>Npar F</u>	<u>Pr(F)</u>
Year	3.151	6.59	<0.001
Month	2.58	9.44	<0.001
Log (Flow)	2.89	16.3	<0.001

**Table 4.2.5:** The significance of each term when included in the final additive semi-parametric model at the River site Linn of Dee.

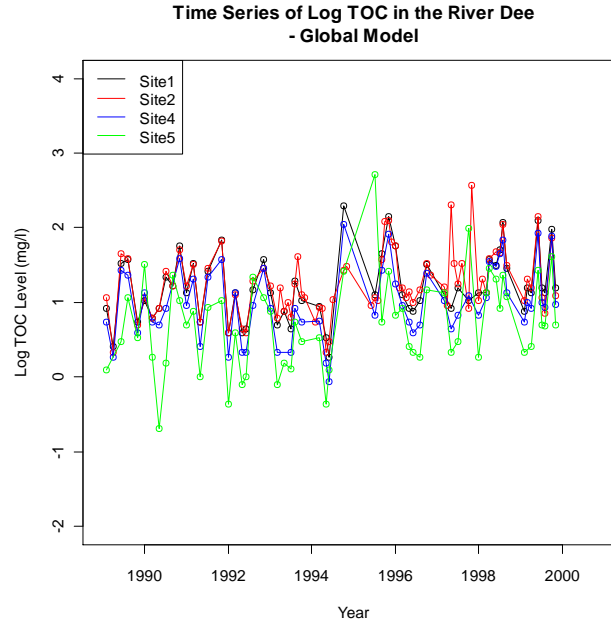
Comparison of Final Linear and Additive Models Fitted to the River Dee Sites Using Approximate F-tests						
Site	Model	RSS	Df	F-statistic	Rejection Region	Preferred Model
Bridge of Dee	Linear	4.865	59	3.6	3.2	Additive
	Additive	4.342	57.09			
Milltimber	Linear	10.173	115	5.4	2.11	Additive
	Additive	7.583	108.248			
Potarch Bridge	Linear	6.622	58	2.18	2.71	Linear
	Additive	5.868	54.78			
Linn of Dee	Linear	18.33	88	5.15	2.37	Additive
	Additive	14.25	83.37			

**Table 4.2.6: Comparison of final linear and additive models fitted to the River Dee sites using an Approximate F-test.**

As discussed in section 3.3.6, an approximate F-test can be used to compare the final linear models fitted at each site, to the final additive model fitted to each site. The results of the approximate F-tests are summarised in Table 4.2.6. The results of the approximate F-tests suggest that an additive model is more appropriate for explaining the levels of log TOC at the River Dee sites: Bridge of Dee, Milltimber and Linn of Dee. This can be expected, based on comparing the adjusted R-squared values from the final linear and additive models. At the Bridge of Dee, the adjusted R-squared value increases from 56.4% to 59.7%; at Milltimber, there is an increase from 56.2% to 65.7%; and at Linn of Dee, there is an increase from 47.3% to 56.7%. But, a linear model is more appropriate to describe the levels of log TOC at the River Dee site Potarch Bridge [which is not entirely surprising, taking into account the adjusted R-squared values of the linear model (48.6%) and the additive model (47.2%)].

## **4.3 Modelling the Levels of Log TOC on the Main Channel: Finding a Global Model**

The focus of section 4.2 was to investigate the sites situated on the main channel and find a model to appropriately describe log TOC levels at each site – independently of each other. However, this section shall differ from the sections previously discussed in the thesis. The main difference being, that for the first time, the sites will not be treated as being independent of each other, as in reality, the sites are located on the same river. It is of interest to find a model to describe the log TOC levels along the River Dee, taking into account that the data were measured at four different sites along river. A Generalized Additive Mixed Model (GAMM) shall be fitted to the four sites. GAMM's are similar to GAM's; but, contain a useful characteristic – they allow the inclusion of a random site effect to be included in the model, to capture the spatial effect. In the previous section, Table 4.1.1 and the time series plots in Figure 4.1.1 (*a*) highlight the differing time periods available for the log TOC data. Since a global model is sought which best describes the log TOC levels in the River Dee, a common time period is required. The time period chosen, is between 1989 and 2000 as the time series plot in figure 4.3.1 displays.



**Figure 4.3.1: Time series of log TOC levels at the 4 River Dee sites**

### 4.3.1 Global Modelling: Generalized Additive Mixed Models (GAMM's)

In the previous sections, additive models have been fitted to each of the sites, independently of one another. GAMM's build on the ideas already explored, with regards to additive modelling. They allow the modeller to include a random effect within the additive model. Including random effects, such as site, allows the introduction of a spatial effect in the model. GAMM's are a combination of GAM's and Linear-Mixed Effects models. GAMM's have the luxury of fitting covariates as smooth or linear terms, but also including a structure which allows for random effects. The GAM's have already been explored. Linear mixed models take the general form:

$$y = X\beta + Zb + \varepsilon, \quad b \sim N(0, \psi_\theta), \quad \varepsilon \sim N(0, \Lambda\sigma^2), \quad (4.3.1.1)$$

where random vector,  $b$ , contains random effects, with zero expected value and covariance matrix  $\psi_\theta$ , with unknown parameters  $\theta$ ;  $Z$  is a model matrix for the random effects.  $\Lambda$  is a

positive definite matrix, of simple structure, which is typically used to model residual autocorrelation: its elements are usually determined by some simple model, with few (or no) unknown parameters. Often  $\Lambda$  is simply the identity matrix. This extension allows the model a more complex stochastic structure than the ordinary linear model, and, in particular, implies that the elements of the response vector,  $y$ , are no longer independent. (Wood, 2006).

Taking into account the four sites, a GAMM model can be fitted to the River Dee, using the *mgcv* package in the statistical software *R*, where the smoothing parameter was selected using cross validation (Wood, 2006). Letting  $y = \log \text{TOC}$ , Year = Year, Month = Month (fitted as a ‘circular’ term), temperature = T, log Alkalinity = A, pH = pH, log Nitrate = N, log Flow = F and log Sulphate = S, a general form of the Generalized Additive Mixed Model can be written as:

$$y_{ij} = \beta_0 + m_1(\text{year}_{ij}) + m_2(\text{month}_{ij}) + m_3(A_{ij}) + m_4(T_{ij}) + m_5(\text{pH}_{ij}) + m_6(N_{ij}) + m_7(S_{ij}) + m_8(\text{Flow}_{ij}) + a_j + \varepsilon_{ij}. \quad (4.3.1.2)$$

Where  $y_{ij}$  is the level of log TOC at the river site  $j = 1, 2, 3, 4$ . Site is included in the GAMM model as a random effect, represented by  $a_j$  in the model. In the GAMM model  $a_j \sim N(0, \sigma_a^2)$  and  $\varepsilon_{ij} \sim N(0, \sigma^2)$ .

Furthermore, GAMM's allow the inclusion of a spatial correlation structure. In Section 3.3.2, the auto-correlation between residuals at each site was separately investigated. Since a global model is being built, the correlation of residuals between sites must be considered. Cross-Correlation Function plots can be used to assess the correlation. The cross-correlation coefficient at lag  $k$  is defined as

$$r_{xy}(k) = \frac{\sum_t (x_{t-k} - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^T (x_t - \bar{x})^2 \sum_{t=1}^T (y_t - \bar{y})^2}}, \quad (4.3.1.3)$$

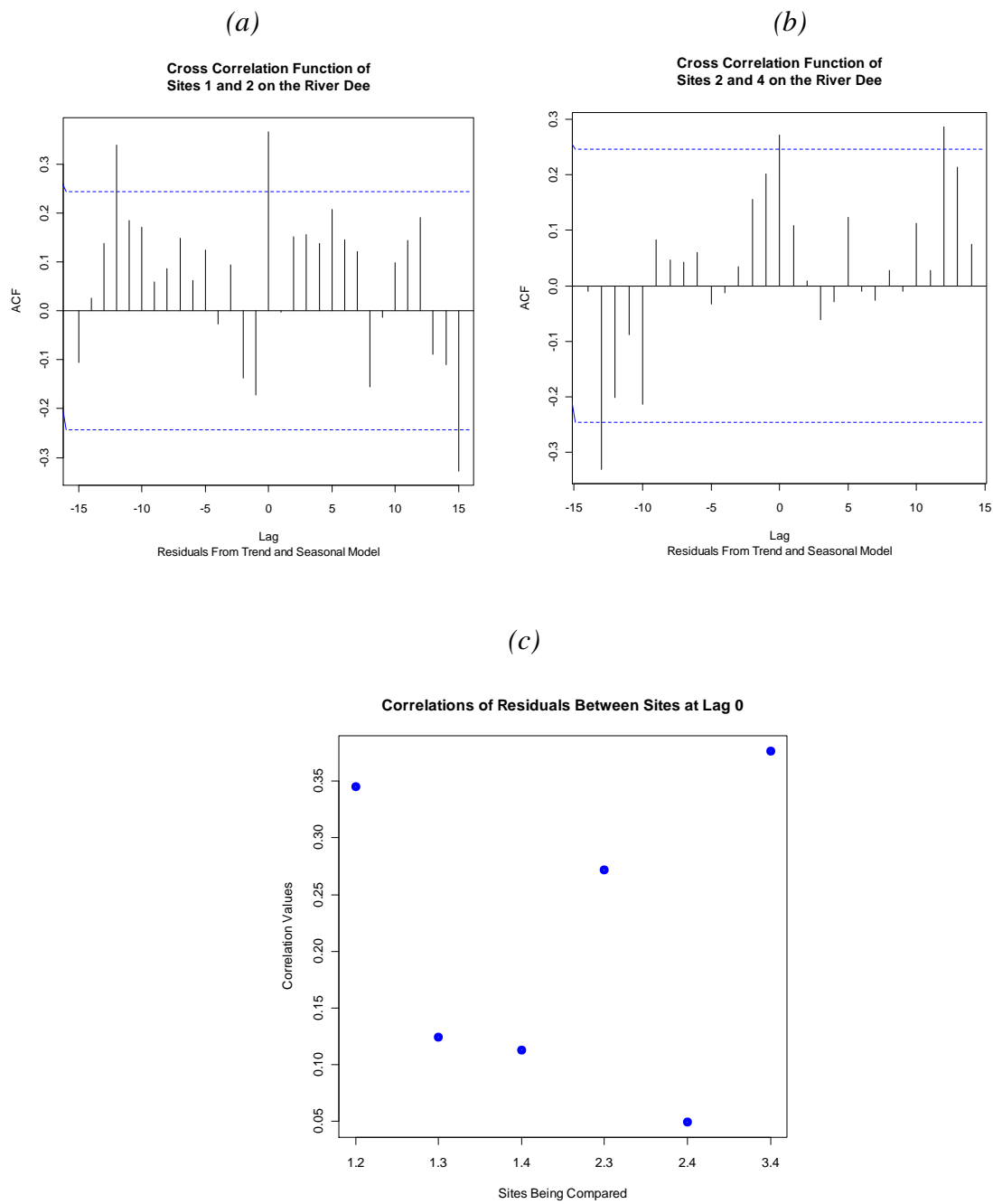
Where the sum in the numerator is computed over all  $t$  for which both  $y_t$  and  $x_{t-k}$  are available;  $r_{xy}(k)$  is a measure of association between values of  $y$  and values of  $x$  that

occurred  $k$  time units previously. The CCF is the collection  $\{r_{xy}(k)\}$  of all sample cross-correlation coefficients. When  $y$  responds to  $x$  after a delay of  $l$  time units (months), the CCF would show a prominent spike at lag  $l$ . (Chandler and Scott, 2011)

Using the residuals from the trend and seasonal linear models fitted to each of the four sites in Section 4.2 (exemplar models displayed in Figure 4.2.1), Cross-Correlation Function plots were constructed. The selection of CCF's plots displayed in Figure 4.3.1.1 [(a) and (b)] highlights the auto-correlation between sites. Furthermore, Figure 4.3.1.1 (c) emphasizes that the lag 0 auto-correlations appear to decrease in an exponential manner as the sites move further apart. The four sites on the River Dee seem to be spatially correlated. There is evidence of inter-station correlation; therefore a correlation structure can be incorporated into the GAMM model. Due to the exponential decrease in lag 0 auto-correlation values discussed earlier, a plausible correlation structure is the exponential:

$$\text{Corr}(\varepsilon_{ts1}, \varepsilon_{ts2}) = \exp(d_{s1,s2}/\psi), \quad (4.3.1.4)$$

where  $d_{s1,s2}$  denotes the Euclidean distance between stations  $s_1$  and  $s_2$  (Chandler and Scott, 2011) and  $\psi$  is the coefficient that explains the strength of the correlation structure as a function of the distance between the sites. The Euclidean distance is the shortest distance calculated between two stations (i.e. the distance calculated if a straight line was to be drawn between the two stations and was measured). Euclidean distance is not the only method of measuring distance between sites. Other approaches have been explored by Cressie et al. (2006) and Ver Hoef et al. (2006), which shall be discussed further in Section 4.4.3. For the purposes of the GAMM model fitted to the sites along the main channel, Euclidean distance shall be used.



**Figure 4.3.1.1: Cross-Correlation Function plots of sites: 1 and 2 (a), 2 and 3 (b), using the residuals from the trend and seasonal linear models fitted in Section 4.2; Plot of lag 0 auto-correlation coefficients from CCF's (c).**



The GAMM model (4.3.1.2) expressed previously was fitted - removing covariates that were not significant at the 5% significance level from the model. Hence, the final GAMM model fitted to the 4 River Dee sites can be expressed as:

$$y_{ij} = \beta_0 + m_1(\text{year}_{ij}) + m_2(\text{month}_{ij}) + \beta_1 A_{ij} + \beta_2 S_{ij} + \beta_3 \text{Flow}_{ij} + a_j + \varepsilon_{ij}. \quad (4.3.1.5)$$

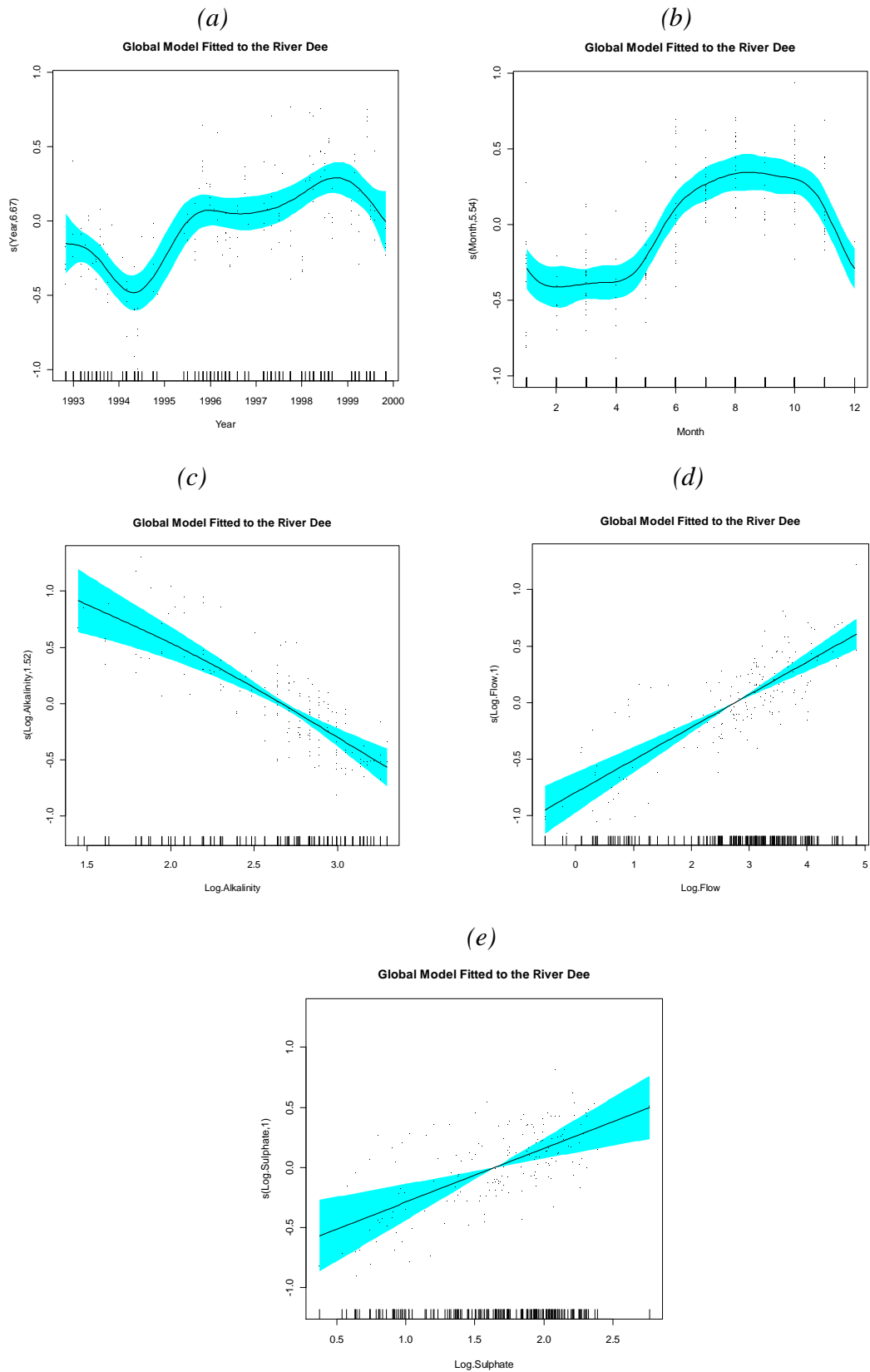
To re-iterate a point made earlier, the error structure in expression (4.3.1.5),  $\varepsilon_{ij}$ , refers to the exponential spatial correlation between sites which is incorporated in the model. Considering Figure 4.3.1, it is evident that the trends of the sites are similar. Therefore, it is no surprise that  $\sigma_a^2$  seen in Table 4.3.1.2 is  $<0.001$  i.e. the standard deviation of the intercept is very low due to the similarity of the trends; hence, there does not appear to be a significant site effect along the main channel.

From the global model fitted to the River Dee, it appears that between the year 1989 and 2000, levels of log TOC have increased gradually, in a non-parametric fashion (emphasized by the smooth curve seen in Figure 4.3.1.2 (a)). As expected, there is a seasonal effect on log TOC (Figure 4.3.1.2 (b)) – levels at their highest during late summer and early autumn. Interestingly, an increase in log Flow and log Sulphate (Figure 4.3.1.2 (d) and (e)), is associated with an increase in log TOC levels; but, an increase in log Alkalinity levels (Figure 4.3.1.2 (c)), is associated with a decrease in log TOC levels. From Table 4.3.1.2 it is evident that log Sulphate is highly correlated with, both, log Flow and log Alkalinity.

The global model, incorporates an exponential spatial correlation structure (expression 4.3.1.4) with  $\psi$  estimated to be 0.238 – hence, there does not appear to be a strong correlation between the sites (as a function of the distance between sites). The final GAMM model seems to be a good fit to the data - the R-squared (Adj) value of 71.9% reinforces this.

The final GAMM model fitted to the 4 sites on the main channel seems to be appropriate; however, if one wishes to consider modeling sites which do not lie on the main channel (i.e. sites which are located on different tributaries which flow into the main channel), including a random site effect in the GAMM model may not be sufficient. A different modeling approach may be more appropriate. A model which takes into account the relationship

between sites located on different tributaries and incorporates a more appropriate way to measure the distance between sites shall be explored in the next section, which shall consider a larger number of sites, over a longer time series, across the river network.



**Figure 4.3.1.2: Year (a), Month (b), Log Alkalinity (c), Log Flow (d) and Log Sulphate (e) effect plots of the GAMM model fitted to the River Dee.**

Summary of Final GAMM Model			
Fitted to the River Dee			
Parametric Coefficients	<u>Estimate</u>	<u>Std. Error</u>	<u>Pr(&gt; t )</u>
Intercept	1.723	0.179	<0.001
Log (Flow)	0.292	0.033	<0.001
Log (Sulphate)	0.422	0.119	<0.001
Log (Alkalinity)	-0.80	0.113	<0.001
Smooth Terms	<u>Npar DF</u>	<u>Npar F</u>	<u>Pr(F)</u>
Year	6.67	16.7	<0.001
Month	5.54	24.02	<0.001

**Table 4.3.1.1: Summary of the Final GAMM model fitted to the River Dee**

Correlation Between Linear Covariates				
	Intercept	Log (Alkalinity)	Log (Flow)	Log (Sulphate)
Year	0.091	-0.064	0.193	-0.073
Log (Sulphate)	0.477	-0.756	-0.606	
Log (Flow)	-0.003	0.098		
Log (Alkalinity)	-0.899			
	Intercept ( $\sigma_a^2$ )		Residuals ( $\sigma^2$ )	
Standard Deviation	<0.001		0.262	

**Table 4.3.1.2: Summary of the Correlations between covariates and standard deviations in the GAMM model fitted to the River Dee.**

## 4.4 The River Dee Network

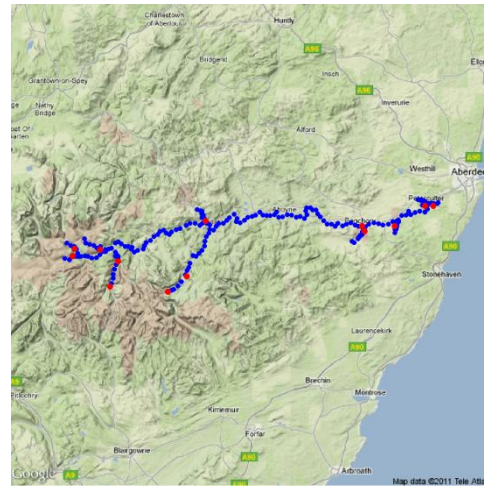
The focus of this section shall be to consider the log TOC levels across the River Dee network (Aberdeenshire) – particularly, tributaries of the network where there are data available. The previous Sub-Sections have focused on modelling the sites situated on the main channel in the network; but, it is of interest to study log TOC levels across a wider region, with the knowledge that not all of the sites are situated on the same stream or channel, and that the sites are not all ‘flow-connected’. This section shall focus on the log TOC levels at 13 River Dee sites (displayed in Figure 4.4.1 in red), with a common time period of 1989 to 2010. The sites under investigation in section 4.4 are listed in Table 4.4.1. Note: some of the sites have missing data across the years, in particular, ‘Banchory Bridge’. As a common period of 1989 to 2010 is now being considered, it was decided to include ‘Banchory Bridge’ for the purposes of analysis - even though it only contributes a small amount of data over the given time period, it still adds to our understanding of the behaviour of log TOC across the network.

A key focus of this section is the comparison between the use of Euclidean distance and river distance to measure the distances between sites. Euclidean distance has been previously explained in section 4.3. River distance is a measurement of the shortest distance between sites following the river. Incorporating these different distance measurements into spatial models has been discussed by Ver Hoef et al. (2006) and Cressie et al. (2006), and as mentioned previously, shall be explored in section 4.4.4.

However, the main aim of this section shall be to conduct spatial modelling over a river network appropriately. In able to achieve this: the behaviour of log TOC shall be studied over space (i.e. across the network); then, the behaviour of log TOC shall be studied over time; before finding an additive model which captures the behaviour of log TOC over time and space appropriately.

(a)

Name	Site
River Dee - Milltimber	1
Culter Burn - Peterculter	2
Sheeoch Burn	3
Water of Feugh – Bridge of Feugh	4
River Dee- Banchory Bridge	5
River Gairn	6
River Muick	7
Dubh Loch – Dubh Loch Outlet	8
River Quoich – Quoich Water	9
Callater Burn	10
Clunie Water – Baddoch Burn	11
River Lui - Lui	12
River Dee – Linn of Dee	13



(b)

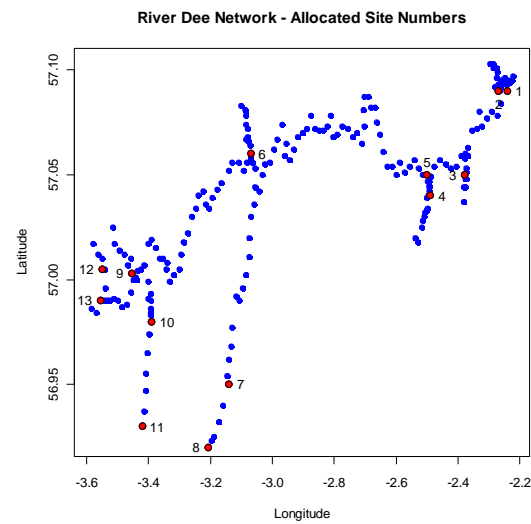


Table 4.4.1

Figure 4.4.1

Table 4.4.1(left): List of the sites under investigation in the River Dee Network.

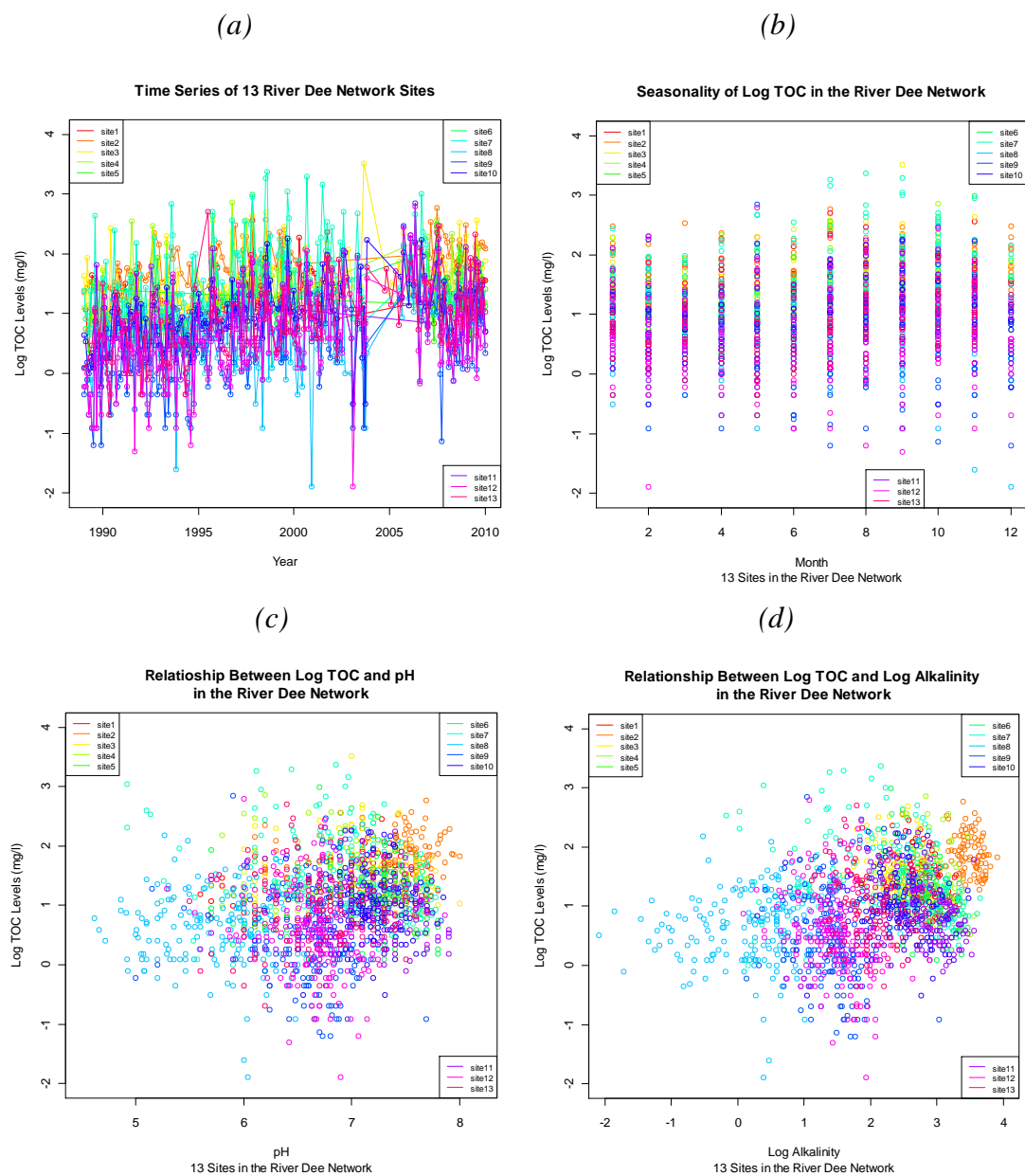
Figure 4.4.1 (right): Portion of the River Dee network plotted in blue and locations of 13 River Dee network sites marked in red on Figure 4.4.1 (a); and the corresponding ‘site number’ of each site is stated on Figure 4.4.1 (b).

### 4.4.1 Trends, Seasonality and Relationships

The thirteen time series under investigation can be explored using standard exploratory analysis techniques. The trend and seasonality of log TOC in each of the sites shall be explored graphically; as well as the relationship between log TOC and the following covariates: log Alkalinity, temperature, pH, log sulphate, log nitrate and log flow [Note: site 5 does not have available data for the covariates log alkalinity, log nitrate, log sulphate and log flow].

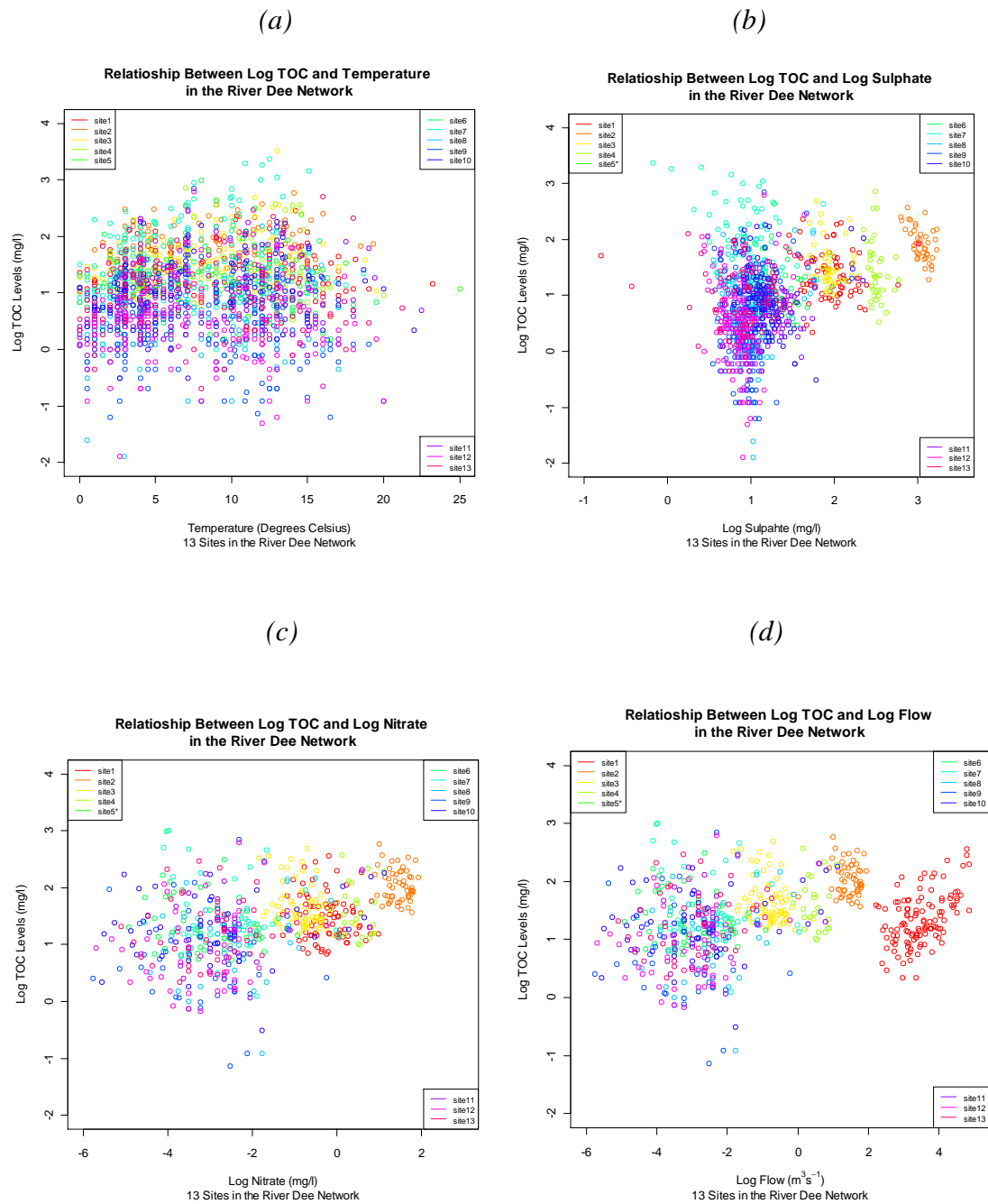
It is of interest to investigate if the thirteen time series in the network are behaving coherently. Figure 4.4.1.1 (a) suggests that most of the log TOC trends in the thirteen time series are similar with the exception of sites 5 and 8, where the trend slightly differs between 2005 and 2010. Furthermore, the seasonal pattern of the log TOC seems to be similar across the network as Figure 4.4.1.1 (b) displays. Overall, the trend and seasonal patterns seem to be similar to the initial impressions expressed earlier in Chapter 2 (with regards to river sites).

It is also of interest, to gain an understanding of the relationship between log TOC and the different covariates at the sites. Inspecting Figures 4.4.1.1 and 4.4.1.2, suggests that the covariates pH and log alkalinity [Figure 4.4.1.1 (c) and (d)] are associated with an increase in the levels of log TOC at the thirteen sites. The other covariates do not seem to have a strong relationship with the log TOC levels.



**Figure 4.4.1.1: The trend (a) and seasonality (b) of log TOC at the thirteen sites; log TOC against pH (c) and log Alkalinity (d) at the thirteen sites.**





**Figure 4.4.1.2: Log TOC against temperature (a), log sulphate (b), log nitrate and log flow (d) at the thirteen sites.**

## 4.4.2 Measures of Spatial Dependence

In section 4.3.1, the spatial correlation was considered for the four sites located on the main channel. Similarly, the spatial correlation of the thirteen River Dee sites scattered across the network (many located on different tributaries), shall be considered. Variograms are used in geo-statistics as a measure of spatial dependence. A variogram is an efficient and effective way of displaying if spatial correlation is, or is not, present. Diblasi and Bowman (2001) developed a test which evaluates the evidence that the empirical variogram changes as a function of  $h$  (where  $h$  represents the distance between locations).

Firstly, if observations are made on a spatial process  $Y(s)$ , where  $s$  denotes a vector of location coordinates, then a key quantity is the variogram, defined by:

$$\gamma(h) = \frac{1}{2} \text{var}\{Y(s+h) - Y(s)\}, \quad (4.4.2.1)$$

where  $h$  denotes a displacement vector. Diblasi and Bowman (2001) states that under the assumption of an intrinsically stationary process, where  $E\{Y(s+h) - Y(s)\} = 0$ , the variogram captures the spatial covariance of the process and is an essential component of any spatial model. A natural estimator is the empirical variogram defined by

$$\hat{\gamma}(h) = \frac{1}{2} \frac{1}{|N(h)|} \sum_{N(h)} \{Y(s_i) - Y(s_j)\}^2, \quad (4.4.2.2)$$

where  $N(h)$  denotes the collection of pairs of observations separated by a distance  $h$ ;  $s_i$  and  $s_j$  denotes different sites (Webster and Oliver, 2001; Diblasi and Bowman, 2001; Hawkins and Cressie, 1984). This test can be used for diagnostic checks for regression models, which need the assumption of independence to hold, and is recommended for examining the variance of residuals from linear models.

Under the assumptions of stationarity and isotropy, a model for the data can be expressed as

$$Y(s) = \mu + \varepsilon(s), \quad (4.4.2.3)$$

where  $\varepsilon(s)$  is normally distributed with mean zero and variogram  $\gamma(h)$ . Under the null hypothesis of independence,  $\gamma(h) = \sigma^2$ . If the errors are independent, the variogram  $\gamma(h)$  is constant; otherwise, there is evidence of spatial correlation. Nonparametric regression is used to create a smooth estimate of the variogram from the difference pairs

$\left( |s_i - s_j|, |Y(s_i) - Y(s_j)|^2 \right)$ , denoted by  $(h_{ij}, d_{ij})$  where  $i < j$ . A smooth estimate of the variogram, can be expressed as:

$$\hat{\gamma}(h) = \sum_{i < j} w_{ij} d_{ij}, \quad (4.4.2.4)$$

where the weights  $w_{ij}$  sum to one and shrink with the distance of  $h_{ij}$  from the point of estimation  $h$ . (Dibiasi and Bowman, 2001)

The *sm* library in the statistical software *R*, allows one to build a variogram, using the test built by Dibiasi and Bowman (2001), which assesses the presence of spatial correlation. The test produces a p-value, of the null hypothesis that  $\gamma(h) = \sigma^2$ .

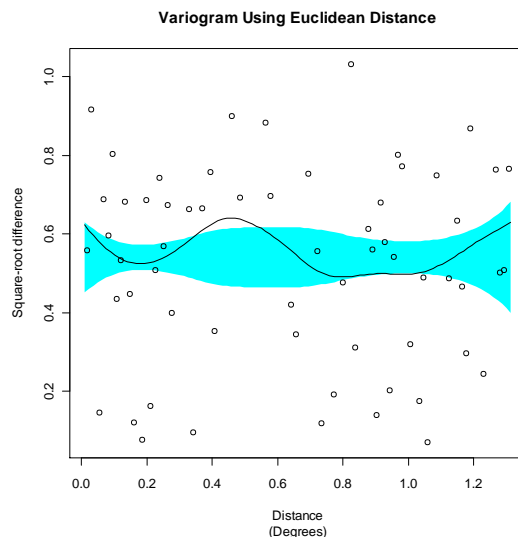
A simple linear regression was carried out for a single time point, which included the thirteen log TOC levels and the location of each site (longitude and latitude). This time point was chosen to be March 2009, as there was data available for all thirteen sites. Hence, the following linear model was fitted, where  $y$  = a vector of 13 log TOC values (one for each site), longitude = Long, and latitude = Lat:

$$y_i = \alpha + \beta_1 \text{Long}_i + \beta_2 \text{Lat}_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2) \quad (4.4.2.5)$$

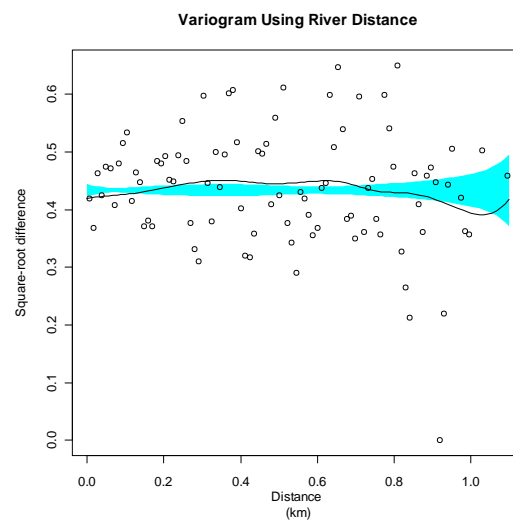
Using the residuals from the fitted linear model 4.4.2.6, a variogram could be constructed using the Euclidean distance as displayed in Figure 4.4.2.1 (a). The distance is measured in degrees, where 1 degree relates to approximately 69.17 km. The test of spatial independence produced a p-value which was equal to 0.736. Therefore, we fail to reject that the log TOC levels at the thirteen locations are spatially independent, based on Euclidean distance being used. Figure 4.4.2.1 (a) displays that the variogram  $\gamma(h)$  seems fairly constant.

The test developed in the *sm* package in the statistical software *R*, was originally designed for the use of Euclidean distance. However, it was possible to construct a variogram using river distance, where the river distance is taken to be the shortest distance between sites following the river path. The spatial coordinates of each site were marked on an ordinance survey map, and then the river distance (*km*) between each site and site 1 was measured. The variogram constructed using river distance is displayed in Figure 4.4.2.1 (*b*). Again, the residuals from the fitted linear model 4.4.2.6 were used. The test of spatial independence using river distance provided a p-value which was equal to 0.881. Again, we fail to reject that log TOC levels at the 13 locations are spatially independent, based on river distance being used. Figure 4.4.2.1 (*b*) displays that the variogram  $\gamma(h)$  seems to be fairly constant. Based on the variogram, spatial dependence does not seem to be an issue, and the conclusions are not altered by the distance measurement used. However, an important point to raise is that the variogram ignores how the river flows between each of the sites i.e. the ‘flow-connectedness’ of the sites is not taken into consideration.

(a)



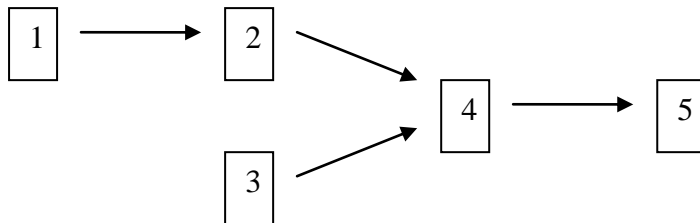
(b)



**Figure 4.4.2.1: Variograms of 13 River Dee network sites, using Euclidean distance (a) and River Distance (b) [lowess curve fitted to both plots].**

### 4.4.3 Flow Connected Sites

Understanding the spatial dependence between sites is important; however, understanding how the water flows between each of the sites is equally important (i.e. are the sites flow connected?). A Directed Acyclic Graph (DAG) (Whittaker, 1990) is an effective method of explaining the meaning of sites being or not being flow connected as Figure 4.4.3.1 displays.



**Figure 4.4.3.1: Directed Acyclic Graph used to express sites which are flow- connected across a network.**

The Directed Acyclic Graph shows: sites 1,2,4 and 5 are flow connected; sites 3, 4 and 5 are flow connected; but, sites 1 and 2 are not flow connected with site 3.

#### 4.4.4 Moving Average Constructions and Valid Covariances

As mentioned previously, appropriately measuring the distance between sites is very important. Choosing the “best” method has been at the heart of current debates and has been thoroughly discussed in recent papers (Ver Hoef et al., 2006; Cressie et al., 2006). Both Cressie et al. (2006) and Ver Hoef et al. (2006) discuss developing valid covariance structures to be incorporated in variograms, when working with river networks.

Ver Hoef et al. (2006) discuss the use of river distances and developing valid spatial autocovariance models for river networks. They argue that the application of typical spatial autocovariance functions based on Euclidean distance may not be valid when using river distance (Ver Hoef et al., 2006). Ver Hoef et al. (2006) use moving average constructions (also called kernel convolutions) to develop suitable models for such networks.

Barry and Ver Hoef (1996) showed that a large class of auto-covariances can be developed by creating random variables as the integrations of a moving-average function over a white noise random process,

$$Z(s) = \int_{-\infty}^{\infty} g(x - s | \theta) W(x) dx, \quad (4.4.4.1)$$

Where  $W(x)$  is a white noise process and  $g(x | \theta)$  is called the moving average function and it is defined on  $\mathcal{R}^1$ . The moving average function can be chosen, but it must have a finite volume in order to create a stationary process. Typically functions centred on 0 are chosen, where most of their mass occurs as well. The moving-average construction allows a valid auto-covariance to be expressed as,

$$C(h | \theta) = \begin{cases} \int_{-\infty}^{\infty} (g(x | \theta))^2 dx + v_j^2 & \text{if } h=0, \\ \int_{-\infty}^{\infty} g(x | \theta) g(x - h | \theta) dx & \text{if } h>0, \end{cases} \quad (4.4.4.2)$$

where it is assumed that the integrals exist and a discontinuity,  $\nu_j^2$  at  $h=0$ , which is the “nugget” effect in geo-statistical terms, is allowed. The moving average construction can be used to build valid models for streams, but also account for water flow. (Ver Hoef et al., 2006)

It is necessary to include in Equation (4.4.4.2) a proper weighting to compensate for the effect in the variance caused by splits in some part of the river (Ver Hoef et al., 2006). The idea is to provide a weight to those cases where there are splits upriver in such a way that the sum of all of them is equal to 1 (Ver Hoef et al., 2006; Rincon, 2009). An appropriate weighting, is to define the sites or the streams making up the network as being flow connected or not. Hence, Equation (4.4.4.2) can be modified to account for proper weighting:

$$C(s_i, t_j | \theta) = \begin{cases} 0 & \text{if } s \text{ and } t \text{ are not flow-connected} \\ \prod_{K \in B_{i,[j]}} \sqrt{\omega_k} C_1(h) & \text{Otherwise.} \end{cases} \quad (4.4.4.3)$$

Where  $C_1(h) = \int_{-\infty}^{\infty} g(x | \theta) g(x-h | \theta) dx$  and recall that  $d(s_i, t_j)$  is the distance between  $s_i$  and  $t_j$  on the river network;  $\omega_k$  is a weight for each stream on the network; and  $K \in B_{i,[j]}$  is the set of all stream sections in the river network, that are between section  $i$  and section  $j$ .

However, Cressie et al. (2006) build on this, putting forward the idea of using both Euclidean distance and river distance. Cressie et al. (2006) use a kernel that is non-negative, linear decreasing, and puts zero weight on river distances larger than  $r$ :

$$K_r(d) = (1 - d/r)I(0 \leq d \leq r). \quad (4.4.4.4)$$



Cressie et al. (2006) express the covariance function as:

$$\text{cov}(Y(s), Y(t)) = \lambda \sigma^2 (\Omega(t) / \Omega(s))^{\frac{1}{2}} \left\{ 1 - \frac{3}{2} \frac{|s-t|}{r_1} + \frac{1}{2} \left( \frac{|s-t|}{r_1} \right)^3 \right\} + (1-\lambda) \sigma^2 \left\{ 1 - \frac{3}{2} \frac{|s-t|}{r_2} + \frac{1}{2} \left( \frac{|s-t|}{r_2} \right)^3 \right\}$$

(4.4.4.5)

The parameter  $\lambda \in [0,1]$  is incorporated to control the amount of spatial dependence described by a river distance in relation to the amount of spatial dependence described by Euclidean distance.

Appropriately capturing the nature of the spatial locations in river networks is very important. The papers by Ver Hoef et al. (2006) and Cressie et al. (2006) highlight that there are different way to tackle this problem; but also highlight, the difficulty of appropriately capturing the relationship between stations located in a river network. In section 4.4.5, a non-parametric technique developed by O'Donnell (2011) [which is based on Ver Hoef's model for a variogram] shall be used to model log TOC over an entire river network.

### 4.4.5 Modelling the River Network

The main aim of Section 4.4 is to build a spatiotemporal model i.e. a model which captures the behaviour of log TOC over time and space in the River Dee network. However, a natural starting point is to consider the behaviour of log TOC over space initially. This sub-section shall explore the log TOC levels across the network. To examine the behaviour of log TOC over space, one time point was chosen – the log TOC values for March of 2009. This particular point in time was chosen as there was log TOC data available for all thirteen sites (note: other time points fitted the criteria and could have been chosen!). Again, further investigation into the use of Euclidean and river distance shall be explored in this sub-section.

To model log TOC over the entire network, a non-parametric technique developed by O'Donnell (2011) can be implemented to capture the behaviour of log TOC. This technique

is based on a very simple local mean smooth function. O'Donnell (2011) adapted Ver Hoef's model for a variogram, so that it could be used as non-parametric smooth, as seen in expression (4.4.5.1). This technique allows one to smooth observations over space. O'Donnell's method shall be carried out to obtain a smooth estimate at each of the 13 known locations; but also, predict smooth estimates at unknown locations across the network. Expression (4.4.5.1) fits a smooth value that is a weighted average of the observations, where the weights are based on the Ver Hoef covariance structure. Expression (4.4.5.1) takes into consideration the distance between locations and whether or not the locations are flow connected. Both, Euclidean and river distance shall be used. Obtaining smooth estimates at the known and unknown locations will provide an indication of the behaviour of log TOC over space.

Firstly, the connectedness between the sites needs to be defined. This is an important step in the modelling of the river network. The connectedness can be expressed in a  $n \times n$  matrix, where  $n$  is the number of sites, there are 1's on the diagonal, and the off-diagonal corresponds to a 1 if the sites are flow-connected and a 0 if they are not. In the River Dee network, a  $13 \times 13$  matrix shall be used.

The distance between sites, shall be defined as the distance from each site to site 1 across the river network, where the river flows towards site 1. The Euclidean distance and the river flow distance (km) between each site and site 1 were calculated.

An estimate for  $\hat{m}(x)$  can be attained using a local mean estimator, using expression (4.4.5.1). The local mean estimator ensures that more weight is given to the observations whose covariate values  $x_i$  lie close to the point of interest  $x$  (Bowman and Azzalini, 1997).

$$\hat{m}(x) = \frac{\sum_{i=1}^n y_i w(d;h) \delta_i(x)}{\sum_{i=1}^n w(d;h) \delta_i(x)} \quad (4.4.5.1)$$

The  $y_i$  refers to the 13 log TOC values at each station. The weight function chosen,  $w(d;h)$ , corresponds to a normal kernel density function centred on zero, with standard deviation  $h$ .

The smoothing parameter  $h$  controls the width of the kernel function, and hence the degree of smoothing applied to the data. As the smoothing parameter increases, the resulting estimator misses some details in the curvature of the data. As the smoothing parameter decreases, the estimator begins to track the data too closely and will end up interpolating the observed points. (Bowman and Azzalini, 1997; Rincon 2009). The distance between point  $x$  and site 1 is defined by  $d$ .  $\delta_i$  denotes

$$\delta_i(x) = \begin{cases} 1 & \text{if point } x \text{ is flow-connected to } x_i, \\ 0 & \text{otherwise} \end{cases} \quad (4.4.5.2)$$

which enables an estimate  $\hat{m}(x)$  to be obtained using only flow connected points in the river network.

To obtain estimates of the log TOC levels across the network: the river distance and Euclidean distance between 217 new locations and site 1 was calculated; and the flow connectedness with the 13 known sites in the network was calculated.

Figure 4.4.5.1 (a-d) displays the smooth estimates of the known and unknown locations using Euclidean distance, with different choices of the smoothing parameter ‘h’ (i.e. h=5, 10, 15, 20). Similarly, Figure 4.4.5.2 (a-d) displays the smoothed estimates using river distance. After exploring the use of different values of ‘h’, 15 seemed to be the most appropriate as it did not over-fit, nor, the contrary. Figure 4.4.5.1 displays that changing the value of ‘h’ seemed to have little effect on the smooth estimates using Euclidean distance.

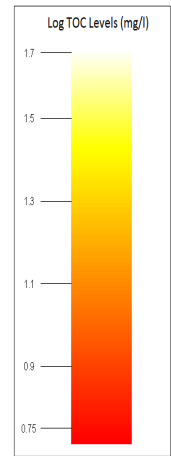
Studying Figure 4.4.5.2 (a-d) suggests that as the river flows downstream towards the sea, the levels of log TOC appear to increase; this is not necessarily clear from Figure 4.4.5.1 (a-d). Furthermore, comparing Figure 4.4.5.1 to Figure 4.4.5.2, the plots suggest that the use of river distance between sites seems to be more appropriate – river distance gives a lower root mean square error value (0.08 compared to 0.31), suggesting that it is a more appropriate distance measurement for river networks.

From Figures 4.4.5.1 and 4.4.5.2 it is apparent, that there is a distinct contrast in the levels of log TOC between sites 7 and 8. A plausible reason for this is not clear from the detail of the map – after reaching the monitoring station, denoted as site 8, the water flows into Dubh Loch, before reaching site 7. This flow-path, could possibly explain the high levels of log TOC in this particular section of the network.

(a)  
Euclidean Distance; h=5



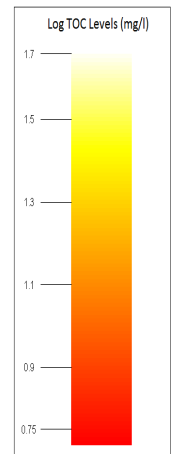
(b)  
Euclidean Distance; h=10



(c)  
Euclidean Distance; h=15



(d)  
Euclidean Distance; h=20

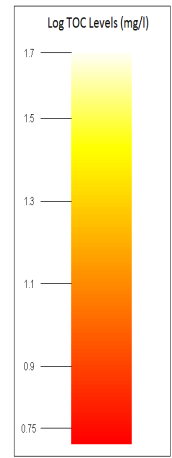


**Figure 4.4.5.1: Smooth estimates of 13 known locations and 217 new locations across the RiverDee Network, using Euclidean distance with the smoothing parameter  $h=5$  (a), 10 (b), 15 (c) and 20 (d). Log TOC values from the month of March in the year 2009 were selected.**

(a)  
River Distance; h=5



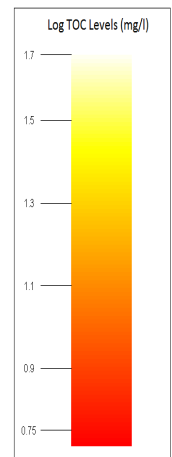
(b)  
River Distance; h=10



(c)  
River Distance; h=15



(d)  
River Distance; h=20



**Figure 4.4.5.2: Estimates of 13 known locations and 217 new locations across the RiverDee Network, using River distance with the smoothing parameter  $h=5$  (a), 10 (b), 15 (c) and 20 (d). Log TOC values from the month of March in the year 2009 were selected.**

## 4.4.6 Visualising Trend Over Space

Having examined the behaviour of log TOC levels over the network at one point in time, the natural next step is to consider the temporal trend of log TOC across the network. An exploratory and effective way to visualise the trend of log TOC over time and space, is to use the same ideas expressed in the previous sub-section; but this time, use four individual points in time. The analysis previously used the log TOC values from March 2009; however, for the purpose of the plots in Figure 4.4.6.1, log TOC levels are used from March, 1990, 1997, 2000 and 2009. Due to the missing data present in site 5, it was not possible to include site 5 in Figure 4.4.6.1 [(b) and (c)] for the years 1997 and 2000. To re-iterate a point expressed earlier, river distance between the sites shall be used as the distance measurement for analysis in the rest of the thesis.

Based on the month of March, inspection of Figure 4.4.6.1 [(a)-(d)] complies with the subjective impressions gained earlier. Levels of log TOC seem to increase between the years 1990 and 2000, particularly where the river rises in the Cairngorms. Comparing Figure 4.4.6.1 (a) to Figure 4.4.6.1 [(b) and (c)] highlights the main increase in log TOC levels between the years 1990 and 2000. The log TOC levels in the sites located where the river rises in the Cairngorms (sites 8-13) are predominantly coloured dark red and orange in (a); but, this is not the case, when the years 1997 and 2000 are considered – the colour scale suggests an increase in this part of the network. The subjective impression gained in earlier chapters is supported further by Figure 4.4.6.1, as it shows that the log TOC levels appear to slightly decrease between the year 2000 (c) and 2009 (d). These plots effectively display the trend over time and space; however, it is important to remember that these plots only consider the month of March across four different years!

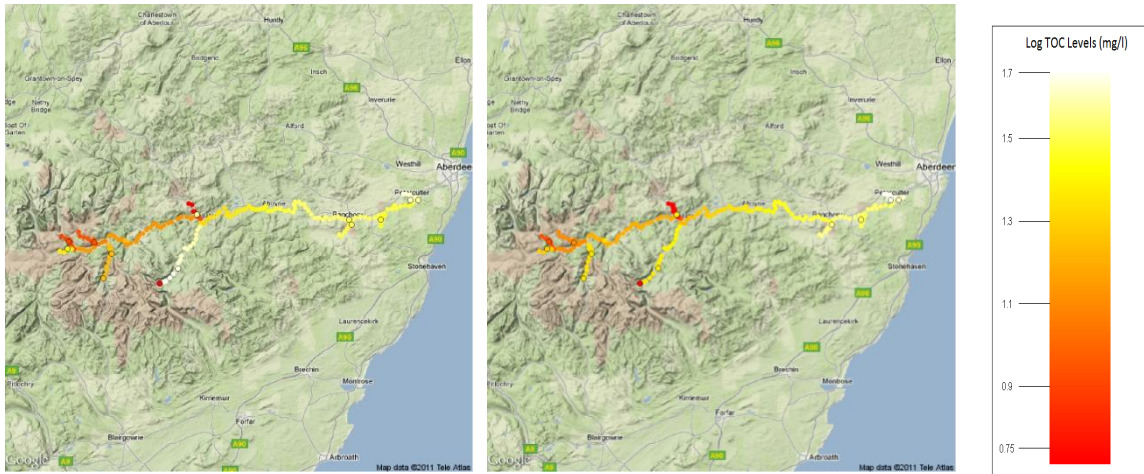
A point of interest is the levels of log TOC found in the stream where site 7 is situated – they are consistently high throughout the years. It is only in 2009 a decrease is seen. The interference of Dubh Loch in the river flow between sites 7 and 8, again, is a plausible explanation.

Even without the inclusion of site 5 in the years 1997 and 2000, it is clear from each of the four points in time, as the river flows through the network, towards sites 1, the levels of log TOC gradually increase.



**1990** (a)

**1997** (b)



**2000** (c)

**2009** (d)



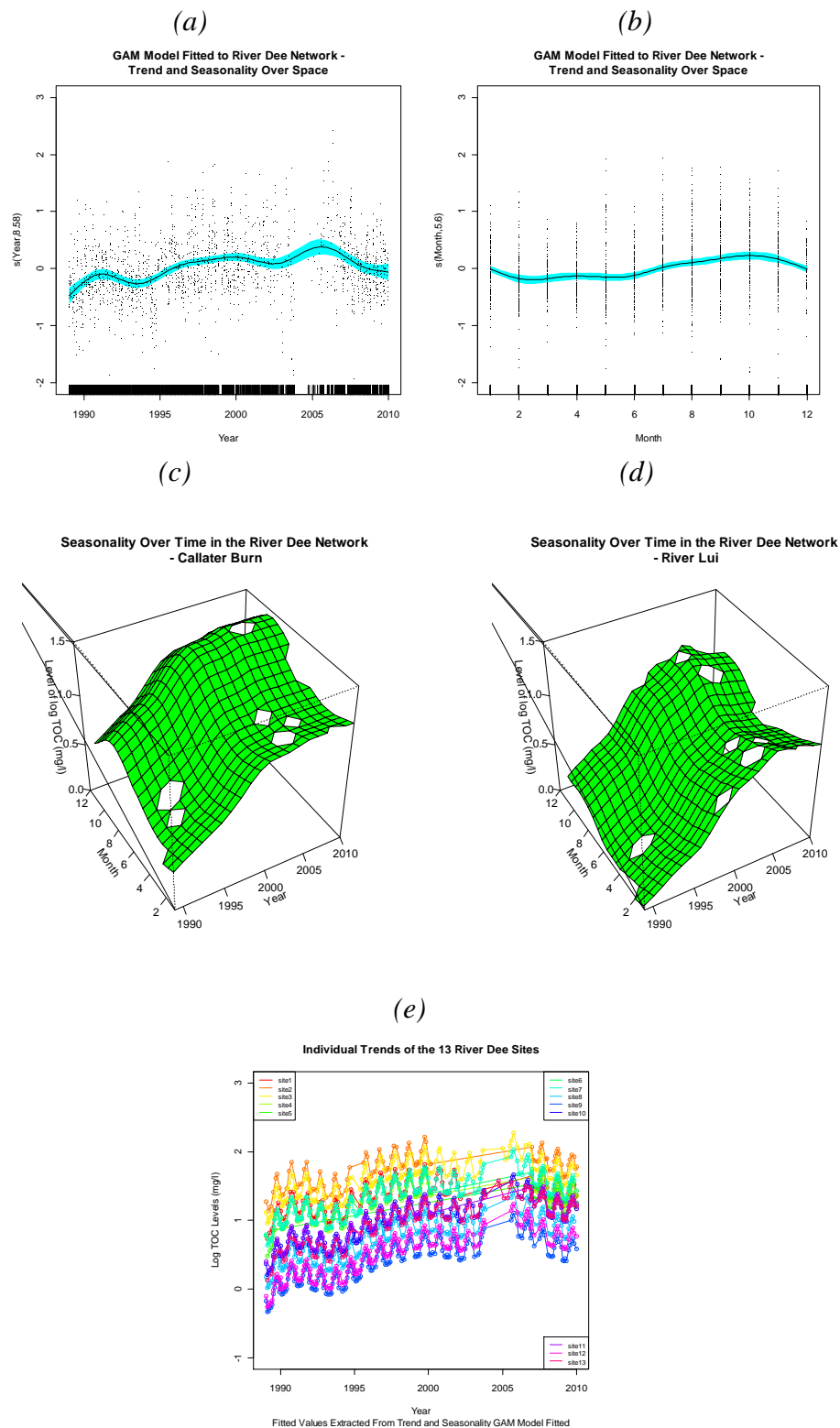
**Figure 4.4.6.1: Estimates of 13 known locations and 217 new locations across the River Dee Network, using river distance (km) and month of March for the years 1990 (a), 1997 (b), 2000 (c) and 2009 (d). Note: site 5 is not included in the years 1997 and 2000.**

#### 4.4.7 Modelling the River Dee Network: Non-Parametric Regression Over Time and Space

After exploring the trend over space graphically, it is now appropriate to move from fitting smoothed log TOC estimates across the network at different points in time, to fitting a model which capture smoothly the behaviour of the log TOC levels between the years 1989 and 2010 across the network. Previously, in section 4.3.1, a GAMM model was fitted to 4 sites situated on the main channel, which included a random site effect. However, since a river network is being considered, including site as a random effect does not seem appropriate. Alternatively, a GAM model can be fitted over time which captures space more effectively. A common way is to include the spatial location of the site as a bivariate term i.e.  $s(\text{longitude}, \text{latitude})$ . Furthermore, a GAM model can be fitted, so that it incorporates a time and space interaction, as it is plausible that the trend in log TOC levels differ between sites.

Initially, a GAM model was fitted to the River Dee sites (still assuming the  $\varepsilon_i$  are independent with mean 0 and constant variance  $\sigma^2$ ), which focused on the trend, seasonality and spatial location of the sites. As it is possible that the log TOC levels may differ between sites, the interactions between ‘year’ and ‘site’, and ‘month’ and ‘site’ are included in the initial GAM model fitted. Letting  $y = \log \text{ TOC levels of the 13 sites}$ ;  $\text{Year} = \text{Year}$ ;  $\text{Month} = \text{Month}$ ;  $\text{Site} = \text{Site Number}$ ,  $\text{Location} = \text{Spatial Location (longitude, latitude)}$ ; the following GAM model can be fitted,

$$y_i = \beta_0 + m_1(\text{Year}_i) + m_2(\text{Month}_i) + m_3(\text{Location}_i) + m_4(\text{Year}_i * \text{Site}_i) + m_5(\text{Month}_i * \text{Site}_i) + \varepsilon_i$$
$$i = 1, \dots, n \quad (4.4.7.1)$$



**Figure 4.4.7.1: Effect plots of the trend and seasonality GAM model fitted to the thirteen sites: Year (a), Month (b). 3D Trend and Seasonality plots of Callater Burn (c) and River Lui (d). Fitted values extracted from GAM model, for each site separately (e).**

Summary of the Trend and Seasonality Additive Model Fitted to the River Dee Network			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	-0.81	0.1	<0.001
Smooth Terms	Npar Df	Npar F	Pr(F)
Year	8.52	13.59	<0.001
Month	3.42	16.74	<0.001
Year:Site	2.0	10.87	<0.001
Location	2.79	30.2	<0.001

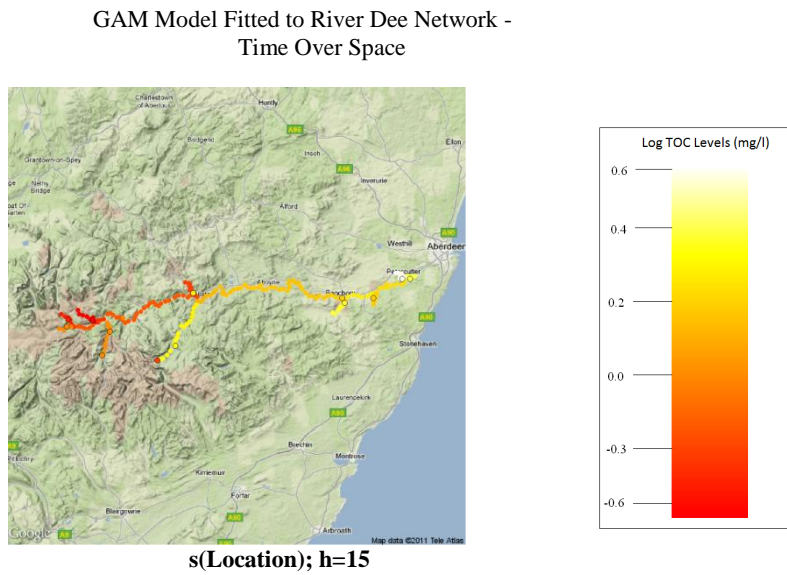
**Table 4.4.7.1: The significance of each term, when included in the trend and seasonality additive model, at the River Dee network.**

Table 4.4.7.1 displays the significant terms included in the trend and seasonality GAM model fitted to the thirteen sites. The effect plots of the additive model can be seen in Figure 4.4.7.1. The initial impression of the trend and seasonality of log TOC coincides with the effect plots displayed in Figure 4.4.7.1 (a) and (b), respectively. Since the interaction term, (Year and Site) is significant, this leads one to believe that the levels of log TOC slightly differ between the thirteen sites across the years. Figure 4.4.7.1 [(c) and (d)] supports this idea – it displays the seasonality over time at two of the network sites (Callater Burn and River Lui). The 3D plots highlight the similarity of the seasonal patterns at each site (which supports the non-significant Month and Site interaction term in the GAM); but, the 3D plots highlight the slight difference in trends, particularly from 2000 onwards (which supports the significant Year and Site interaction term in the GAM). Table 4.4.7.1 highlights, that the term ‘location’ is significant. This suggests that the spatial location of the site within the network will have an effect on the log TOC levels, which coincides with idea that as the river

flows downstream towards site 1, the levels of log TOC seem to increase. Figure 4.4.7.1 (e) displays the fitted values extracted from the trend and seasonality GAM model for each of the thirteen sites – inspection of this plot, would suggest that it is likely, that groups of sites in the network are behaving coherently. It seems plausible, that groups of sites in the network share a common trend, specifically, sites located near each other.

To investigate whether the term  $m_3(\text{Location}_i)$  was capturing the trend over space appropriately, the partial residuals of the term were calculated (Rincon, 2009). To attain the partial residuals of the term  $m_3(\text{Location}_i)$ ,  $r_i$  was calculated in the following manner:

$$r_i = y_i - \hat{y} - \hat{m}_1(\text{Year}_i) - \hat{m}_2(\text{Month}_i) - \hat{m}_4(\text{Year}_i * \text{Site}_i) \quad i = 1, \dots, n \quad (4.4.7.2)$$



**Figure 4.4.7.2: Smoothed mean partial residuals of the term  $m_3(\text{Location}_i)$  for each of the thirteen sites (a); but, also the smoothed partial residuals of the other 217 new locations.**

Once  $r_i$  had been calculated, the mean partial residuals of each site could then be calculated. Figure 4.4.7.2 displays the smoothed mean partial residuals of the term  $m_3(Location_i)$  for each of the thirteen sites; but, also the smoothed partial residuals of the other 217 new locations. The average partial residuals provide a guide to the pattern (Rincon, 2009). These values have a black outline around their circles on the plot. Figure 4.4.7.2 suggests that the term  $m_3(Location_i)$  is capturing the trend over space suitably.

The adjusted R-squared value (37.7%) suggests that the trend and seasonality additive model could be possibly improved by the inclusion of covariates. Letting  $y = \log$  TOC levels of the 13 sites; Year = Year; Month = Month; Site = Site Number, Location = Spatial Location (longitude, latitude); T = temperature; A = log alkalinity; pH = pH; S = log sulphate; N = log nitrate; and F = log flow, the following GAM model can be fitted,

$$y_i = \beta_0 + m_1(Year_i) + m_2(Month_i) + m_3(T_i) + m_4(A_i) + m_5(pH_i) + m_6(S_i) + m_7(N_i) + m_8(F_i) + m_9(Location_i) + m_{10}(Year_i * Site_i) + \varepsilon_i$$

$$i = 1, \dots, n \quad (4.4.7.3)$$

Again, terms that were not significant at the 5% level were removed from the GAM model, and the model was refitted. Hence, the final GAM model fitted to the thirteen sites can be expressed as:

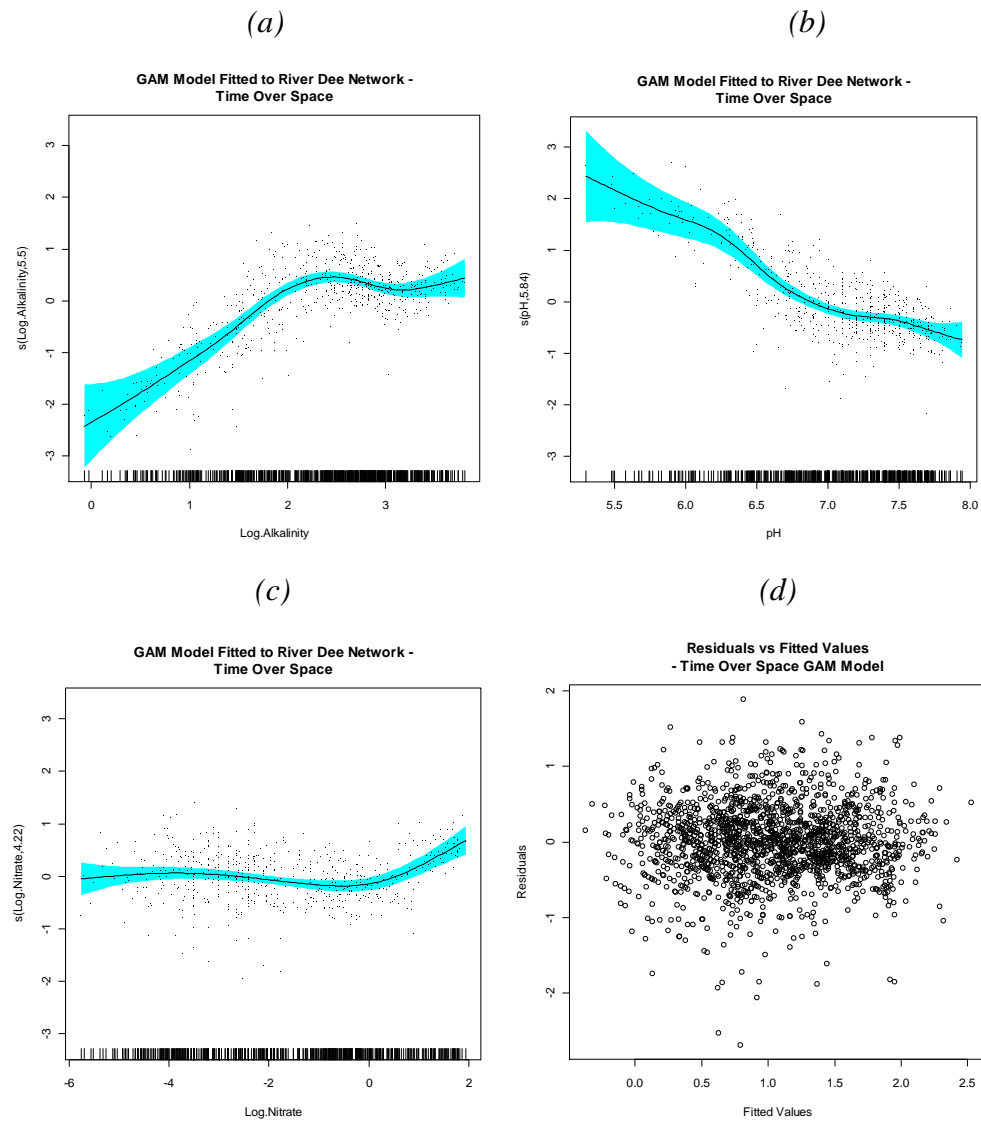
$$y_i = \beta_0 + m_1(Year_i) + m_2(Month_i) + m_3(A_i) + m_4(pH_i) + m_5(N_i) + m_6(Location_i) + m_7(Year_i * Site_i) + \varepsilon_i$$

$$i = 1, \dots, n \quad (4.4.7.4)$$

The final GAM model fitted to the River Dee network is summarized in Table 4.4.7.2 and the effect plots of log alkalinity, pH, and log nitrate corresponding to the model can be seen in Figure 4.4.7.2 (a),(b) and (c) respectively.

Summary of the final GAM fitted to the River Dee Network			
Parametric Coefficients	Estimate	Std. Error	Pr(> t )
Intercept	1.22	0.15	<0.001
Smooth Terms	Npar Df	Npar F	Pr(F)
Year	7.86	7.04	<0.001
Month	5.43	8.65	<0.001
Log Alkalinity	5.5	16.62	<0.001
pH	5.84	13.8	<0.001
Log Nitrate	4.22	8.86	<0.001
Location	2.08	6.4	<0.001
Year : Site	2.0	6.81	<0.001

**Table 4.4.7.2: The significance of each term, when included in the final additive model, at the River Dee network.**



**Figure 4.4.7.3: Effect plots of the final GAM model fitted to the thirteen sites: Log Alkalinity (a), pH (b) and log nitrate (c). Residuals vs Fitted values from the final GAM model fitted to the thirteen sites (c).**



Figure 4.4.7.3 (a) suggests that as the levels of log alkalinity increase in the River Dee network, the levels of log TOC appear to increase; however, as the levels of log alkalinity increase above approximately 2.5, the levels of log TOC seem to decrease. With regards to pH, Figure 4.4.7.3 (b) suggests that an increase in the pH level in the River Dee network is associated with a smooth decrease in log TOC levels. Figure 4.4.7.3 (c) suggests that log nitrate levels below 0 are not associated with any change in log TOC levels; but, log nitrate levels above 0 are associated with a sharp increase in log TOC levels. The residuals vs fitted values plot in Figure 4.4.7.3 (d) suggests that there is no issues with the model fitted to the data; having said this, even though the adjusted R-squared value (45%) suggests that it is an improvement on the trend and seasonality additive model, it is not a reasonable fit to the data. Future work on the River Dee network, could explore the inclusion of other covariates to improve the final model fitted.

## **4.4.8 Conclusions of the River Dee Network**

At first, Sections 4.1 to 4.3 considered the five sites located on the main channel of the River Dee independently of one another. The exploratory analysis suggested that there was a common signal – the log TOC levels were increasing steadily until the early 2000's, which was followed by a weaker increase in the remaining years; there was a seasonal pattern evident in all sites; and the covariate 'log flow' seemed to influence log TOC levels at all sites (where flow data was available). Two modelling approaches were explored – the use of linear models and additive models. The approximate F-tests used concluded: additive modelling was appropriate at three of the sites; and a linear model was more appropriate at Potarch Bridge. [Noting – analysis of Banchory Bridge was deemed not to be of any value in Sections 4.2 and 4.3, due to the large amount of missing data].

A Generalized Additive Mixed Model was then fitted to capture the common signals of the four sites located on the River Dee itself. The final GAMM model (4.3.1.5) revealed that there was a significant trend and seasonal pattern amongst the four sites; but also, the covariates log Alkalinity, log Sulphate and log Flow influenced log TOC levels at the four sites. The final GAMM model (4.3.1.5) had advantages over the linear and additive models fitted to the sites individually: a global model was found to describe the behaviour of log TOC along the River Dee; it allowed the inclusion of a random site effect; and a spatial correlation structure (exponential) could be incorporated in the model to account for the correlation between sites (as a function of the distance between sites). In a sense, the spatial correlation structure highlighted, that the distance between sites along the river, had an influence on the levels of log TOC.

Having considered sites which were located on the main channel in the River Dee network, it was then of interest to consider sites located on streams and estuaries which flowed into the main channel. To gain an understanding of the behaviour of log TOC across the network, a non-parametric smoothing technique developed by O'Donnell (2011) was chosen. O'Donnell's smoothing technique effectively captured the structure of the River Dee network – the distance between sites, and how each site was 'flow connected' were taken into consideration, allowing smooth log TOC estimates for the known and unknown locations in the network to be obtained. Initially, the behaviour of log TOC was studied over space – particularly, the log TOC values during March 2009. As this chapter was interested in comparing Euclidean to river distance as an appropriate distance measurement between sites, O'Donnell's non-parametric smoothing technique was conducted using both measurements. Regardless of which distance measurement was used, it was clear, that as the river flows through the network, downstream towards site 1, the levels of log TOC seem to increase. Based on the visual inspection of plots, and comparison of the root mean square error values, it was concluded, that river distance seems to be a more appropriate measurement between sites and was used in subsequent analysis. A natural progression from investigating the behaviour of log TOC over space was to monitor the trend of log TOC over time and space. To achieve this, four points in time were chosen – the log TOC values from March 1990, 1997, 2000 and 2009. Unfortunately, 'site 5' was missing data in the years 1997 and 2000. The trend appeared to coincide with initial impressions previously formed in earlier sections

– in the month of March (in the chosen years) the log TOC levels seemed to increase throughout the 1990's up until the early 2000's, and then “level off”. Levels of log TOC seem to increase between the years 1990 and 2000, particularly at the sites located where the river rises (near the Cairngorms). The plots in Figure 4.4.6.1 are effective for visualising the trend over space in the network; however, these plots only considered four time points, and log TOC values from the month of March. Without time restraints, it would have been interesting to study the behaviour of log TOC over time and space, for the four seasons, and also, a greater number of years. However, the missing data may cause problems, as finding months which had log TOC data for every year and every site was a challenge.

The GAMM model appropriately captured the behaviour of sites situated on the same channel; however, in order to capture the common signals of the sites located across the network, a different approach was required. A GAM was fitted (4.4.7.1) to initially capture the trend and seasonality of the log TOC levels across the network. The spatial location was included in the model as a covariate to capture the space element of the network; and the interactions between ‘year’ and ‘site’, and ‘month’ and ‘site’, were included, as it was thought that the levels of log TOC may be differ between sites. Having fitted the GAM model, it was clear that the trend, seasonality, spatial location and interaction between the site and year were all significant. The inclusion of the spatial location in the model, effectively capture the spatial element of the network - based on Figure 4.4.7.2, the inclusion of the smooth term, ‘location’, seemed to capture the trend over space suitably (using river distance and including flow connectedness).

Plotting the fitted values from the trend and seasonality GAM model [Figure 4.4.7.1 (*e*)] for each of the thirteen sites, supported the significant interaction terms in the model and the idea that the log TOC levels differ between sites; but leads one to believe, that it was more plausible that groups of sites in the network were behaving coherently, particularly, sites located near each other. It seems plausible, that groups of sites in the network share a common trend.

The trend and seasonality GAM model was improved by the inclusion of the covariates log alkalinity, pH and log nitrate. As the levels of log alkalinity increase in the River Dee

network, the levels of log TOC appear to increase; however, as the levels of log alkalinity increase above approximately 2.5, the levels of log TOC seem to decrease. With regards to pH, it seems that as the pH levels increase in the network, the log TOC levels seem to decrease in a smooth, gradual, manner.

In the River Dee network, it appears that log nitrate levels below 0 are not associated with any change in log TOC levels; but, log nitrate levels above 0 are associated with a sharp increase in log TOC levels. The additive model including covariates did improve the trend and seasonality model; however, the adjusted R-squared value was only 0.45. It is possible that data for other environmental covariates could be explored and used to explain the behaviour of log TOC in the River Dee network.

This chapter has focussed on the log TOC levels of sites located in the River Dee network, finding an appropriate model for a site or a group of sites. However, it is of interest to explore the coherency of log TOC levels at different sites – are the log TOC levels at sites located close to each other behaving similarly? The next chapter shall explore different techniques of measuring coherency; and consider log TOC levels of sites located in regions of Scotland.

# Chapter 5

## Coherency

Identifying common signals and trends across monitoring stations in Scotland is the key focus of this thesis. In other words, we are measuring the coherency of sites. Coherency is the main theme throughout this thesis. The Cambridge Dictionary of Statistics defines coherency to be:

"In time series analysis, it is used to describe the strength of association between two or more time series where the possible dependence between the series is not limited to simultaneous values but may include leading, lagged and smoothed relationships."

Measuring coherency allows an investigation into whether the behaviour of log TOC is similar across the rivers and lochs in Scotland. It also allows an insight into whether sites with similar climatic factors, biological processes or geographical surroundings are coherently similar. Coherency has been used in a variety of statistical genres to identify 'common signals'. This section, shall explore how different authors have approached measuring 'coherency'. A literature review has been conducted in order to obtain an understanding of the variety of ways in which different papers have tackled the problem of measuring coherency. Following the literature review, this chapter applies several methods (Seasonal Mann Kendall and Dynamic Factor Analysis) of measuring coherency to the River

Dee network and a selection of Scottish regions and compares the results with the analysis in Chapter 4.

## 5.1 Literature Review

Correlation has been used as a measure of coherence and temporal coherence. To obtain an initial idea of the coherency between different time series, Munoz-Carpena et al., (2005) proposed calculating cross-correlations between all response and explanatory variables, across all time series. The cross-correlation coefficients are a useful exploratory tool and also provide a measure of the relationship between paired data sets; but, do not capture the simultaneous interactions of multivariate time series.

Many papers have estimated coherence by calculating the correlation between time series for each of the different variables (Magnuson et al., 1990; George et al., 2000; Magnuson et al., 2006b; Pace et al., 2002; Patoine et al., 2006; Benson et al., 2000). The mean correlation and the percentage of strong correlations are calculated for each pair of time series across all variables, and for each variable across all time series pairs. Magnuson et al. (1990) state, ‘temporal coherence, which we define as the degree to which different locations within a region behave similarly through time, is a useful concept because the more coherent different locations are, the easier it is to generalize about specific regional responses to variation in climatic factors, changes in land use, or even environmental stress from contaminants’. Magnuson et al., (1990) calculate the arithmetic mean correlation ( $\bar{r}$ ) for each variable and each time series pair. Then the percentage of strong correlations was calculated for each variable and each time series pair – the percentage of correlation coefficients larger than a threshold of +0.67 [which is the critical value for a one-tail test of the correlation coefficient for significance at the 0.05 level with 5 degrees of freedom]. This procedure was carried out with the view that strong correlations represented the strength of temporal coherence.

Baines et al., (2000) measure the coherency of physical and chemical properties of different lakes in Wisconsin, by simply fitting a linear regression which ‘predicts observations of a

variable in one lake, against simultaneous observations of the same variable in another' (Baines et al., 2000; Bloch et al., 2010) and then comparing the  $r^2$  values. The  $r^2$  value has the advantage of simple interpretation, where  $r^2$  is the proportion of the variance in the response variable that can be explained by the model. Baines et al., (2000) use the 'Pearson product-moment correlation coefficient,  $r$ , to inspect distributions of correlations.'

Ghanbari et al., (2011) use methods which are based on the linear spectral approach used by Ghanbari et al., (2009) to analyze coherence between time series. In a linear spectral approach, Hanson et al., (2004), compute the spectrum of each time series, and then the spectra of the two time series are compared to find common frequency bands in their variability. This differs from a coherency function approach, where 'the co-spectra and cross spectra are computed and these functions are used to calculate the squared coherency that objectively shows the frequency bands that are common between two time series' (Ghanbari et al., 2011). Ghanbari et al., (2011) estimate the squared coherency, in a similar manner to Jenkins et al., (1968) and Bloomfield (1976). 'The values of coherency estimates were considered significant at the 95% level of confidence when they were larger than the critical value  $T$  derived from the upper 5% point of the  $F$ -distribution on  $(2, d-2)$  degrees of freedom, where  $d$  is the degrees of freedom associated with the univariate spectrum estimates' (Ghanbari et al., 2011). Cygnus Research International (CRI) argue that calculating the coherency function, is an alternative, and more effective measurement of coherency among time series, than the use of correlation coefficients.' The CRI, state that the coherency function 'is a function of frequency' and therefore, it has the ability to 'show at which frequencies two sets of time series data are coherent and at which frequencies they are not'.

Curtis et al., (2005) recently measured coherency in a medical sense. The focal point of their paper was not based on 'which parts of the brain are active during working memory delays, but instead on what might persistent activity represent'. Coherence is formally used to characterize functional interactions between different regions of the brain. Curtis et al., (2005) think of the coherence statistic, 'as a correlation in frequency space'. Where the coherence between time series is defined by:

$$Coh_{xy}(\lambda) = \frac{|f_{xy}(\lambda)|^2}{[f_{xx}(\lambda)f_{yy}(\lambda)]}$$

‘Where  $f_{xy}(k)$  is the cross-spectrum of  $x$  and  $y$ , and  $f_{xx}(k)$  is the power spectrum of  $x$  (Brillinger, 2001; Muller et al., 2001). It is a normalized measure from 0 to 1, where 0 indicates an absence of any linear relation, and 1 indicates that the signals are perfectly related by a linear magnitude and phase transform’. (Curtis et al., 2005)

Lange et al., (2004) measure coherency using the cross correlation function between ‘runoff and global-long time indices’. The cross-correlation function measures the correlation between two time series at  $n$  different time lags; and Lange et al., (2004) fit 95% significance bands to their cross-correlation function plot, to provide an insight into the significance of each correlation coefficient. Lange et al., (2004) admit that ‘additional methods are required to elaborate further on the linkages between the filtered time series and plausible drivers’. However, Lange et al., (2004) managed to successfully identify the ‘synchronous behaviour of the signals confined to a geographical region’.

To determine whether or not a large number of time series are behaving coherently, Blenckner et al., (2007) used a method more commonly used in biostatistics: meta-analysis. Many biostatistical papers have used and discussed meta-analysis (Marshall et al., 1996; Fine et al., 1993); but, Blenckner et al., (2007) have used meta-analysis to measure the common signals in lakes across Europe. Coherency can be measured through the use of meta-analysis techniques. They are very effective when one wishes to investigate whether a large number of sites behave coherently – a meta-analysis compares results from numerous studies, providing an aggregated statistical test which is more powerful than statistical tests performed on the sites individually. The meta-analysis can provide information on the overall magnitude of an effect, on whether that effect differs among contrasting categories of studies, and how the variation is distributed within and among categories (similar to analysis of variance). Furthermore, meta-analysis allows the factors that influence the overall pattern of coherence to be determined, and offers the additional advantage of allowing each individual study to be weighted by the number of samples included in the study. The meta-analysis is not flawed with respect to outliers, hence, possible effect sizes are not due to outliers from



one site. The overall effect size ( $E^{++}$ ) and the corresponding 95% confidence intervals (CI) is calculated for all target variables as outlined by Rosenberg et al. (2000).

$$E^{++} = \frac{\sum_{i=1}^n w_i \times E_i}{\sum_{i=1}^n w_i},$$

where  $E_i$  is the calculated effect size for the  $i^{\text{th}}$  study. The variance of  $E^{++}$  is the reciprocal of the sum of the weights given to each of the  $n$  studies:

$$S_{E^{++}}^2 = \frac{1}{\sum_{i=1}^n w_i}.$$

The confidence interval (CI) of  $E^{++}$  is then given by

$$CI = E^{++} \pm t_{\alpha/2[n-1]} \times S_{E^{++}},$$

where  $t_{\alpha/2[n-1]}$  is the two-tailed value of Student's  $t$ -distribution at the critical level  $\alpha$ , and  $n$  is the number of individual studies. An overall effect is considered to be significant if the CI does not include zero (Gurevitch et al., 2000).

Folster et al., (2005) considered the coherency of an even larger number of time series than Blenckner et al. (2007). Folster et al., (2005) aimed to investigate the common signals of 80 lakes in Sweden. Similar to Magnuson et al. (1990), pearson product moment correlation coefficients ( $r$ ) were calculated for each variable, for every lake pair. Again, the  $r$  value is a measure of coherence between a lake pair, with regards to that particular variable. As the lakes were widely spread across Sweden, the dependence of coherence on distance between lakes was studied by linear regression. In order to investigate if the coherency of a lake-pair was related to the similarity of the traits of two lakes, Folster et al., (2005) calculated the relative difference,  $D_x$ , for a number of lake and catchments traits.

$$D_x = \frac{|X_a - X_b|}{X_a + X_b}$$

where  $x_a$  and  $x_b$  are the characteristics for lakes a and b.  $D_x$  was calculated for select variables. Folster et al., (2005) then explored the relationship between coherency and  $D_x$  graphically and by linear regression.

A more efficient and effective way to capture the coherency of multivariate time series, was developed by Zuur et al., (2003) – a technique known as Dynamic Factor analysis. Dynamic Factor analysis has been used to identify common signals of time series and specify the number of common trends present in multivariate time series in a variety of papers (Zuur et al., 2003a; Zuur et al., 2004; Zuur et al., 2003b; Munoz-Carpena et al., 2005). Munoz-Carpena et al., (2005) state that the aim of Dynamic Factor analysis is to ‘choose the smallest number of common trends as possible – because, although increasing the numbers of common trends leads to a better model fit, it results in more information that needs to be interpreted’ (which can often be difficult).

The results from DFA are interpreted in terms of the estimated parameters, the canonical correlations, and match between model estimates and observed values. The goodness-of-fit of the model can be assessed by visual inspection, the coefficient of efficiency (Nash et al., 1970) and Akaike’s Information Criterion (Akaike., 1974; Munoz-Carpena et al., 2005). Choosing the “best” Dynamic Factor model, to describe the  $n$  time series, takes into account all of these factors.

Zuur et al., (2003a) also discuss another criticism, that DFA is based on normality. As DFA can be seen as a regression model, and therefore relies on the same underlying assumptions, then non-normality does not prove to be an issue. Similar to linear regression, if there is a problem of non-normality due to outliers, different transformations of the data can be performed to achieve normality.

Seasonality within time series is a key issue when using DFA. If the time series has cyclic or seasonal components present in the data, they will be masked and included in the trend component of the Dynamic Factor model. Zuur et al., (2004) discuss this issue: when analyzing seasonal data, the most common time series models fitted, takes the following form:

$$Y(t) = T(t) + S(t) + I(t) + e_t$$

Where  $Y(t)$  is a univariate time series,  $T(t)$  is the trend,  $S(t)$  the time-dependent seasonal component,  $I(t)$  can contain cycles, explanatory variables or autoregressive terms and  $e(t)$  is the error. Dynamic Factor Analysis, in a sense, is a multivariate extension of this model. However, Zuur et al., (2003a) states that such a Dynamic Factor model results in computational problems. Zuur et al., (2004), suggest an alternative method of dealing with the seasonal component i.e. remove the seasonal component from the data before any analysis (de-seasonalising). Removing the seasonal component can be executed through calculating the monthly averages over all the years, and then simply subtracting the appropriate average from each value. Another possibility, suggested by Harvey et al., (1989) is to model the monthly data as a parametric cosine function. If seasonality seems fairly constant over time, these are plausible methods. These approaches assume that there is no shift in seasonal maxima or minima. If there was evidence of a shift in maxima or minima, the strategy would have to be re-considered to allow for this fluctuation.

Alternatively, literature addressing Seasonal Dynamic Factor analysis has been published recently, by Alonso et al., (2011). Alonso et al., (2011) apply seasonal dynamic factor analysis (SeaDFA) techniques to electricity market forecasting – the SeaDFA allows the extraction of the common factors of a vector of time series, and the estimation of a seasonal multiplicative Vector Auto Regressive Integrated Moving Average (VARIMA) model, so that both regular and seasonal dynamics can be modelled.

A Bayesian approach has been employed by statisticians to measure the coherency of time series (Lopes et al., 2008; Strickland et al., 2009), based on the use of dynamic factor analysis methods to develop methods which consider spatial dynamic factor analysis. Strickland et al., (2009) argues ‘data sets that vary across space and time have become so large that “standard” approaches are no longer feasible’. Lopes et al. (2008) and Strickland et al., (2009) believe that Bayesian methods are the most appropriate method for performing dynamic factor analysis and dealing with seasonal or cyclic components. Lopes et al., (2008) explain that ‘the temporal dependence is modelled by latent factors while the spatial dependence is modelled by the factor loadings; the spatial dependence is incorporated into the factor loadings by a combination of deterministic and stochastic elements; the number of factors is treated as another unknown parameter and fully Bayesian inference is performed via a reversible jump Markov Chain Monte Carlo Algorithm’.

Nye et al., (2008) measure coherency in three different ways: dynamic factor analysis (as previously discussed); loess smoothing; and minimum/maximum autocorrelation factor analysis (MAFA). Nye et al., (2008) fit the ‘locally weighted regression smoother (loess) to sections of data by weighting points relative to their distance from the target value’, where the ‘smoothed mean trend of all survey time-series is simply the average of the smoothed trends calculated for each time-series’. Nye et al., (2008) also used MAFA, which is a ‘data reduction technique’, similar to principal components analysis (Zuur et al., 2007). MAFA takes into account, the temporal autocorrelation structure, which is used to detect the number of statistically significant trends. Once the statistically significant trends have been identified, ‘the canonical correlations between extracted trends and both individual time series and explanatory variables’ are calculated. (Nye et al., 2008)

Another appropriate technique to measure coherency, is a non-parametric test known as the Mann Kendall. It is a method used for trend analysis, predominantly in an environmental setting (Gilbert., 1987; Chen et al., 2008; Esterby., 1993; Mann., 1945; Kendall., 1975; Weyhenmeyer., 2008). The Mann Kendall test has appealing characteristics: missing values do not cause any problems; and the data do not need to follow a particular distribution. The Mann Kendall test is used to determine whether or not there is a trend within a particular time series, and an estimate of the slope is calculated using a Sen estimator (Sen., 1968b). To assess coherency, the Mann Kendall test can be performed on numerous stations (i.e. a number of time series), and the homogeneity of the stations can be measured (Gilbert., 1987) – this allows us to infer if there is a common signal at a group of sites.

Gilbert (1987) addresses the issue of seasonal cycles present in data, and discusses the seasonal Kendall test developed by Hirsch et al., (1982) – a test built on the fundamentals of the Mann Kendall. The seasonal Kendall test (Hirsch et al., 1982) provides: a slope estimator of the  $i^{\text{th}}$  season for the  $k^{\text{th}}$  year; a test of the homogeneity of trends in different seasons [a test closely related to the procedure developed by van Belle et al., (1984)]; and a test for global trends (van belle et al., 1984).

However, Bloch et al., (2010) highlight that the Mann Kendall test does not ‘reveal how coherent temporal variations, in particular seasonal variations, are between lakes’ (i.e.  $n$

number of time series). Bloch et al., (2010) measure coherency using Kendall's  $\tau$  test, which 'gives a rank correlation coefficient (Kendall's  $\tau$ , ranging from -1 to 1), expressing how good temporal variations of a variable in one lake is following the temporal variations of the same variable in another lake (Helsel and Hirsch, 1992).' Bloch et al., (2010) goes on to explain that using both tests, is useful, as 'the results of both the Kendall's  $\tau$  test and the Mann–Kendall give information about individual variables and how they behave in different lakes.'

Pryor et al., (2009) investigate the coherency of 'century-long precipitation records from stations in the contiguous USA'. Pryor et al., (2009) admits that trend analysis is most easily accomplished by ordinary least squares regression, which has been used extensively in previous studies, has some flaws e.g. it is not 'robust to outliers or to deviations from normality such as might reasonably be expected to characterize "extreme" descriptors'. Pryor et al., (2009) use two different methods: Kendall's tau-based slope estimator (Alexander et al., 2011; Sen, 1968) similar to the papers discussed previously (Gilbert., 1987; Chen et al., 2009; Esterby, 1993; Mann, 1945; Kendall, 1975); and 'application of bootstrap re-sampling (Lunneborg, 2000) of the residuals from OLSR analysis.' 'These residuals are computed and then randomly selected using a bootstrapping technique and added onto the linear fit line from the trend analysis and the trend is re-estimated (Kiktev et al., 2003). This procedure is repeated 1000 times to generate 1000 plausible trends for each station. The trend terms from those 1000 samples are then tested to determine if a zero trend falls within the middle 900 values in an ordered sequence of the distribution of 1000 realizations. If so the original trend is deemed not significant at the 90% confidence level. The trend magnitude is given by the median value of the 1000 samples.' (Pryor et al., 2009). Pryor et al., (2009) found that 'bootstrap techniques generally resolve a larger number of significant trends'.

Potamias et al., (2001) express the importance of coherency: 'Measuring similarity between objects is a crucial issue in many data retrieval and data mining applications'. The main aim of measuring coherency, is to achieve a final outcome, which includes 'the clustering of time series into similar-groups' (Potamias et al., 2001). To achieve a clustering of time series, Potamias et al., (2001), follow a piecewise linear segmentation approach, where the different time-to-time changes, based on their significance according to the full time series, are

weighted. Computing the distances between time series, ‘feeds an appropriate distance-based clustering algorithm in order to form clusters of similar time series’ and uses the ‘neighbour joining clustering algorithm’ by Saitou et al., (1987) to produce clear and informative phylogeny trees and dendograms.

Another plausible method of measuring coherency is to use the “wavelet coherency” method (Sanderson et al., 2010; Hassan et al., 2009; Polansky et al., 2010; Torrence et al., 1998). Wavelets can be used to model the dependence between two non-stationary time series. Polansky et al., (2010) suggest ‘frequency and time–frequency domain methods, embodied by Fourier and wavelet transforms’ as a suitable measurement of coherency – where the use of continuous wavelet transforms, solve some of the limitations of the Fourier analysis. ‘The wavelet transform uses short windows for higher frequencies, which leads to more natural localization in time and scale’ (Sanderson et al., 2010). Torrence et al., (1998) argue that ‘decomposing a time series into time–frequency space, one is able to determine both the dominant modes of variability and how those modes vary in time’, which makes it a very appealing strategy for measuring coherency. The concept of the wavelet cross-spectrum, in terms of the continuous wavelet transform, was introduced by Hudgins et al. (1993), and has since been applied to fields including climatology (Maraun and Kurths, 2004) and neuroscience (Lachaux et al., 2002).

Carey et al., (2010) use Principal Components Analysis (PCA) to measure coherency of time series. Carey et al., (2010) portray PCA to be ‘exploratory in nature’, but goes on to explain that PCA has been ‘previously used to map catchments into similar groupings based on hydrological and other indices, and can provide additional insight when exploring the dependency among factors’ (Pfister et al., 2000; Monk et al., 2007; Tetzlaff and Soulsby, 2008; Carey et al., 2010).

Kent et al., (2007) tackle coherency in a different manner, by performing correspondence analysis of bacterial communities. The Bray-Curtis similarity coefficient (Legendre and Legendre, 1998) is calculated for each sample obtained by Kent et al., (2007) to ‘assess the degree of similarity between bacterial communities obtained from different samples’, using the following:

$$S_{jk} = 1 - \sum \frac{|y_{ij} - y_{ik}|}{(y_{ij} + y_{ik})}$$

where  $y_{ij}$  is the normalized peak area of the  $i^{\text{th}}$  population in the  $j^{\text{th}}$  sample and  $y_{ik}$  is the normalized peak area of the  $i^{\text{th}}$  population in the  $k^{\text{th}}$  sample. Kent et al., (2007) generate a similarity matrix for all possible pairs of samples; this similarity matrix was used to produce an analysis of similarity (ANOSIM) statistics (Clarke and Green, 1988) to test the hypothesis that bacterial communities from the same lake were more coherent than communities in different lakes. Kent et al., (2007) produce a test statistic,  $R$ , for the analysis of similarity [an approach also used by Gremberghe et al., (2007)]. Where, the magnitude of  $R$  provides an indication of the ‘degree of separation between groups of samples, with a score of 1 indicating complete separation and 0 indicating no separation.’

The coherency between sites is at the centre of this thesis. This section highlighted that coherency is measured in many fields and with the aid of different techniques; but, coherency is always assessed with a common aim - to identify common signals.

## 5.2 Methodology

The literature review has highlighted the vast number of techniques which have been used in different papers, to measure coherency. Based on the literature, it seems appropriate that the Seasonal Mann Kendall test and Dynamic Factor analysis shall be applied to the River Dee network (previously explored) and the Scottish regions to gain an idea of the coherency present between the sites.

### 5.2.1 Seasonal Mann Kendall Test

One approach to measuring the coherency of sites is to use the non-parametric Mann-Kendall test for trend (Mann, 1945; Kendall, 1975). Since, the exploratory analysis highlighted the presence of seasonality in log TOC, it is appropriate to use the Seasonal Mann

Kendall (Hirsch et al., 1982; 1982; Van Bell et al., 1984) test to measure the homogeneity of sites. Van Belle and Hughes (1984) developed a test to identify global trends, when using the Seasonal Mann Kendall test. At each station, the data has been split into 4 seasons (Winter, Spring, Summer and Autumn). [It is important to mention that the test could of been applied to months also, instead of simply just the seasons].

The first step, is to compute the Mann-Kendall statistic for each season at each station, in the following manner (where  $S_i$  denotes the statistic computed for season i)

$$S_i = \sum_{k=1}^{n_i-1} \sum_{l=K+1}^{n_i} \text{sgn}(x_{il} - x_{ik}) \quad (5.2.1.1)$$

Now, letting  $S_{im}$  denote the Mann-Kendall statistic for the  $k^{\text{th}}$  season at the  $m^{\text{th}}$  station:

$$Z_{im} = \frac{S_{im}}{[\text{VAR}(S_{im})]^{1/2}} \quad i = 1, 2, 3, 4 \quad m = 1, 2, \dots, n \quad (5.2.1.2)$$

Where  $\text{VAR}(S_{im})$  is obtained by calculating

$$\begin{aligned} \text{VAR}(S_i) = & \frac{1}{18} \left[ n_i(n_i - 1)(2n_i + 5) - \sum_{p=1}^{gi} t_{ip}(t_{ip} - 1)(2t_{ip} + 5) - \sum_{q=1}^h u_{iq}(u_{iq} - 1)(2u_{iq} + 5) \right] \\ & + \frac{\sum_{p=1}^{gi} t_{ip}(t_{ip} - 1)(t_{ip} - 2) \sum_{q=1}^{hi} u_{iq}(u_{iq} - 1)(u_{iq} - 2)}{9n_i(n_i - 1)(n_i - 2)} + \frac{\sum_{p=1}^{gi} t_{ip}(t_{ip} - 1) \sum_{q=1}^{hi} u_{iq}(u_{iq} - 1)}{2n_i(n_i - 1)} \end{aligned} \quad (5.2.1.3)$$



And where  $g_i$  is the number of groups of tied (equal valued) data in season  $i$ ,  $t_{ip}$  is the number of tied data in the  $p^{\text{th}}$  group for season  $i$ ,  $h_i$  is the number of sampling times (or time periods) in season  $i$  that contain multiple data, and  $u_{iq}$  is the number of multiple data in the  $q^{\text{th}}$  time period in season  $i$ . (Gilbert, 1987)

Van Belle and Hughes (1984) then suggest computing the mean over the  $n$  stations for the  $i^{\text{th}}$  season in the following manner:

$$\bar{Z}_{i.} = \frac{1}{n} \sum_{m=1}^n Z_{im} \quad i = 1, 2, 3, 4 \quad (5.2.1.4)$$

And then the mean over 4 seasons for the  $m^{\text{th}}$  station, in the following way:

$$\bar{Z}_{.m} = \frac{1}{4} \sum_{i=1}^4 Z_{im} \quad m = 1, 2, \dots, n \quad (5.2.1.5)$$

And also, the mean over all KM stations and seasons, like so

$$\bar{Z}_{..} = \frac{1}{4 \times n} \sum_{i=1}^4 \sum_{m=1}^n Z_{im} \quad (5.2.1.6)$$

Bearing this in mind, Chi-Square Statistics can be computed and referred to the appropriate corresponding degrees of freedom to test for station and seasonal heterogeneity, as Table 5.2.1.1 displays.

Seasonal Mann Kendall Test	
Chi-Square Statistics	Degrees of Freedom
$\chi^2_{station} = K \sum_{m=1}^M \bar{Z}_{.m}^2 - KM\bar{Z}_{..}^2$	$M - 1$
$\chi^2_{season} = M \sum_{i=1}^K \bar{Z}_{i.}^2 - KM\bar{Z}_{..}^2$	$K - 1$

**Table 5.2.1.1: Summary of Seasonal Mann Kendall Test chi-square statistics and corresponding degrees of freedom.**

## 5.2.2 Dynamic Factor Analysis

Dynamic Factor Analysis (DFA) is a method which has the ability to model the common signals of a group of time series. DFA is a method which can estimate the common trends, effects of explanatory variables and interactions in multivariate time series datasets. The main aim of DFA is to estimate underlying common trends. Therefore, letting the vector  $Y_t = (Y_1, \dots, Y_n)'$  contain the values at year  $t$  for the  $n$  sites, the simplest DFA model contains only one common trend and is given by

$$Y_t = Az_t + \varepsilon_t. \quad (5.2.2.1)$$

Where the elements of  $A$  are called factor loadings and indicate which common trends are important for which of the  $N$  response variables;  $z_t$  represents one common trend at time  $t$ ; the term  $\varepsilon_t$  represents noise components and it is assumed that  $\varepsilon_t$  are normally distributed with expectation 0 and covariance matrix  $R$  (covariance matrix  $R$  is described in more detail later).

To model time lags, the simple Dynamic Factor model (5.2.2.1) can be easily extended to (Zuur et al, 2003a):

$$y_t = A_0 z_t + A_1 z_{t-1} + A_2 z_{t-2} + \dots + A_L z_{t-L} + \varepsilon_t \quad (5.2.2.2)$$

In these models, the response variables are modelled as a function of latent variables at time  $t$ , plus a time delay in these variables (by latent, we mean hypothetical or made up). DFA falls under criticism for being a “latent variable model”: DFA generates latent variables, suggesting that these variables are an existing quantity and can be measured – which, logically thinking, is not always the case. However, if the latent variables represent a factor, e.g., temperature, such a model would be plausible.

If  $A$  is a vector of dimension  $n \times 1$  with unknown loadings, and  $z_t$  is the trend, then

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{n,t} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} z_t + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{n,t} \end{pmatrix} \quad (5.2.2.3)$$

The model with one common trend assumes that all the  $n$  time series follow the same pattern, namely that of  $z_t$ . To obtain the fitted value for each time series, we multiply the trend  $z_t$  by a loading. If the loading is relatively large and positive, we know that the corresponding time series follows the pattern of the trend. If the loading is close to zero, we know it does not follow this pattern. A loading that is relatively large and negative indicates that the time series follows the opposite pattern of the trend. These statements assume that the spread in the  $n$  time series is the same. One way to ensure this is normalisation of the time series prior to analysis. It is also an option to include an intercept:

$$Y_t = c + A z_t + \varepsilon_t \quad (5.2.2.4)$$

The DFA model can be extended to include covariates, similar to that of a linear model. For example:

$$Y_t = c + AZ_t + \beta_0 X_t + \varepsilon_t \quad (5.2.2.5)$$

Where  $X_t$  is the value of the explanatory variable at time  $t$ ; and  $\beta_0$  is a regression coefficient.

The dynamic factor analysis model so far has only considered a group of time series with one common trend – an advantage of DFA, is that it allows for one to  $p$  common trends if groups of time series show similar trends. Equation (5.2.2.6) shows the DFA model extended to two trends:

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{n,t} \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \\ \vdots & \vdots \\ a_{1,n} & a_{2,n} \end{pmatrix} \begin{pmatrix} z_{1,t} \\ z_{2,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{n,t} \end{pmatrix} \quad (5.2.2.6)$$

One can add even more trends, but just as in PCA the interpretation of three or more axes (trends) becomes difficult. (Zuur et al., 2007).

Zuur et al., (2003a) also discuss another criticism, that DFA is based on normality. As DFA can be seen as a regression model, and therefore relies on the same underlying assumptions, then non-normality does not seem to be an issue. Similar to linear regression, if there is a problem of non-normality due to outliers, different transformations of the data can be performed to achieve normality. Furthermore, it is important to note that missing log TOC values do not present a problem when fitting a DFA model.

Previous chapters have highlighted the seasonal pattern present in the time series. Therefore, as suggested by Zuur et al., (2004), the seasonal component is removed from the data before any analysis, through calculating the monthly averages over all the years, and then simply subtracting the appropriate average from each log TOC value (for each time series independently). Therefore, the DFA models fitted will not account for seasonality!

When a DFA model is fitted, it can be modelled using a diagonal error covariance matrix or a non-diagonal error covariance matrix – previously described as ‘covariance matrix R’. A diagonal error covariance matrix indicates the amount of information that cannot be explained by the common trends. A non-diagonal error covariance matrix is similar to that of a diagonal error covariance matrix, but, if present, 2-way interactions between the time series are modelled by the off-diagonal elements. (Zuur, 2011). DFA models can be fitted with both types of error covariance matrices and compared.

Dynamic Factor Analysis aims to find a model with the lowest number of common trends, but still finding a reasonable fit. Akaike’s Information Criterion (AIC) is a measure for goodness of fit which can be used to compare DFA models and choose the “best” model.

## **5.3 Applications of Methodology: River Dee Network**

Having studied the River Dee network in depth in the previous chapter, it seemed appropriate to apply the methods discussed in Section 5.2 to the thirteen River Dee sites. Applying the methods discussed in Section 5.2 shall provide an insight into the coherency of the sites located in the River Dee network.

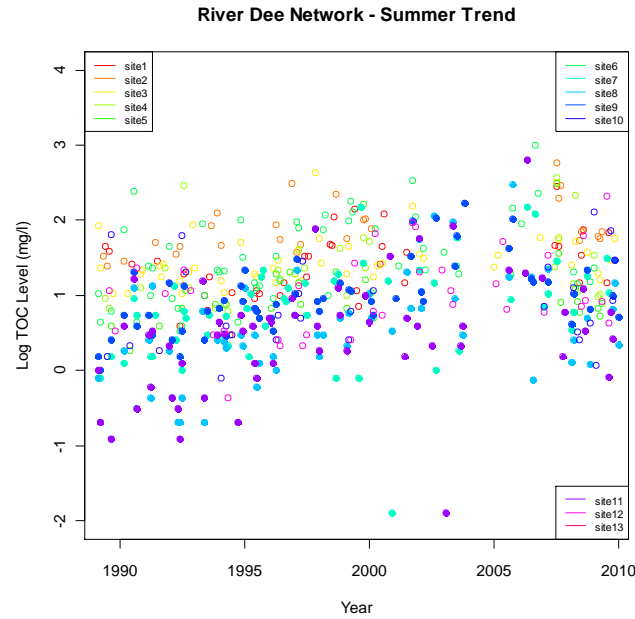
### **5.3.1 Applying the Seasonal Mann-Kendall Test to the River Dee Network**

The heterogeneity of the River Dee sites can be measured using the Seasonal Mann-Kendall test as discussed in Section 5.2.1. The chi-square statistic in Table 5.2.1.1 ( $\chi^2_{station}$ ) tests the null hypothesis that the trend at each site is in the same direction. When applied to the thirteen River Dee sites, the  $\chi^2_{station}$  statistic was equal to 17.55. This value was subsequently referred to the appropriate degrees of freedom stated in Table 5.2.1.1 [where

$\alpha = 0.05$ ] - the chi-square critical value was equal to 19.68. Since  $\chi^2_{station}$  did not exceed this critical value, we fail to reject the null hypothesis, and conclude that the trend at each site is in the same direction in the River Dee network. Similarly, the chi-square statistic in Table 5.2.1.1 ( $\chi^2_{season}$ ) tests the null hypothesis that the trend is in the same direction in each season. Again,  $\chi^2_{season}$  was referred to the appropriate degrees of freedom, as outlined in Table 5.2.1.1. Since  $\chi^2_{season}$  was equal to 33.9 and the chi-square critical value was equal to 7.8, the null hypothesis was rejected. Since  $\chi^2_{season}$  is significant, but,  $\chi^2_{station}$  is not, this means that the trends have significantly different directions in a different season or seasons, but not at different stations. Since this is the case, van Belle and Hughes (1984) developed a chi-square statistic to test the null hypothesis that there was a different trend direction in each season by computing the K seasonal statistics:

$$M\bar{Z}_i^2 \quad i = 1, 2, 3, 4 \text{ seasons} \quad (5.3.1.1)$$

The seasonal statistics for winter, spring, summer and autumn were equal to 0.36, 2.99, 55.63 and 2.21, respectively. These values were referred to a chi square distribution, with 1 degree of freedom, which was equal to 3.84. Hence, the null hypothesis could not be rejected for winter, spring or autumn. But, the null hypothesis was rejected for summer. For all stations, the trend is in the same direction in winter, spring and autumn; but, the trend is not in the same direction during the summer as the test statistic (55.63) is greater than the chi-square critical value (3.84). To conclude, the overall trend is the same at all sites as are the winter, spring and autumn trends, but the summer trend varies between sites.



**Figure 5.3.1.1: The summer trend of the 13 River Dee network sites log TOC values.**

Figure 5.3.1.1 highlights that the summer trend is not in the same direction at all stations – the points of stations 7, 8, 9 and 11 are in bold to emphasize that their log TOC values appear to steadily increase between 1990 and early 2000's before levelling off; compared to the log TOC values at the other stations which remain fairly flat between 1990 and 2010. The season in the River Dee network seems to have a strong influence on the trend.

### 5.3.2 Applying Dynamic Factor Analysis to the River Dee Network

To measure the coherency of the thirteen sites in the River Dee network further, DFA models were fitted using *Brodgar* 2.7.2 (Zuur, 2011). The DFA models were fitted to the 13 time series over the same length of time period: 336 months. Missing log TOC values do not present a problem when fitting a DFA model. At first, DFA models were fitted which only considered common trends using expression (5.2.2.4). DFA models were fitted with varying number of common trends, and either incorporating a diagonal or non-diagonal error covariance matrix. Each time, the AIC value was recorded. Based on the DFA models which only considered trend, Table 5.3.2.1 highlights that a model fitted with two common trends and a non-diagonal error covariance matrix has the lowest AIC value, and is the most appropriate (highlighted in red in Table 5.3.2.1). This suggests that there are two underlying common trends in the River Dee network.

However, covariates can be added to the DFA models to try to explain what is driving the observed trends. Hence, using expression (5.2.2.5), DFA models were fitted with varying numbers of common trends, a diagonal or non-diagonal error covariance matrix and a combination of covariates. Unlike previous chapters, the explanatory variables need to be included in the DFA model as a covariate which is common to all sites. Hence, data from the Met Office has been used. Data on the annual mean temperature (degrees Celsius) and annual rainfall (mm) has been extracted for use in the DFA models, as these explanatory variables are common to all sites and are physical factors thought to influence organic carbon levels (Freeman et al., 2001a; Worrall et al., 2004; Moxley, 2010). The Met Office provides summaries of these explanatory variables for the north, east and west of Scotland – therefore, the appropriate data are used, depending on the location of the sites i.e. data for the east of Scotland is used for the River Dee sites. Again, the AIC value of each DFA model was recorded and is displayed in Table 5.3.2.1.

Table 5.3.2.1 suggest that including covariates, has improved the DFA models. Based on the AIC values, Table 5.3.2.1 suggests that a DFA model with 2 common trends, which includes a non-diagonal error covariance matrix and both explanatory variables (mean

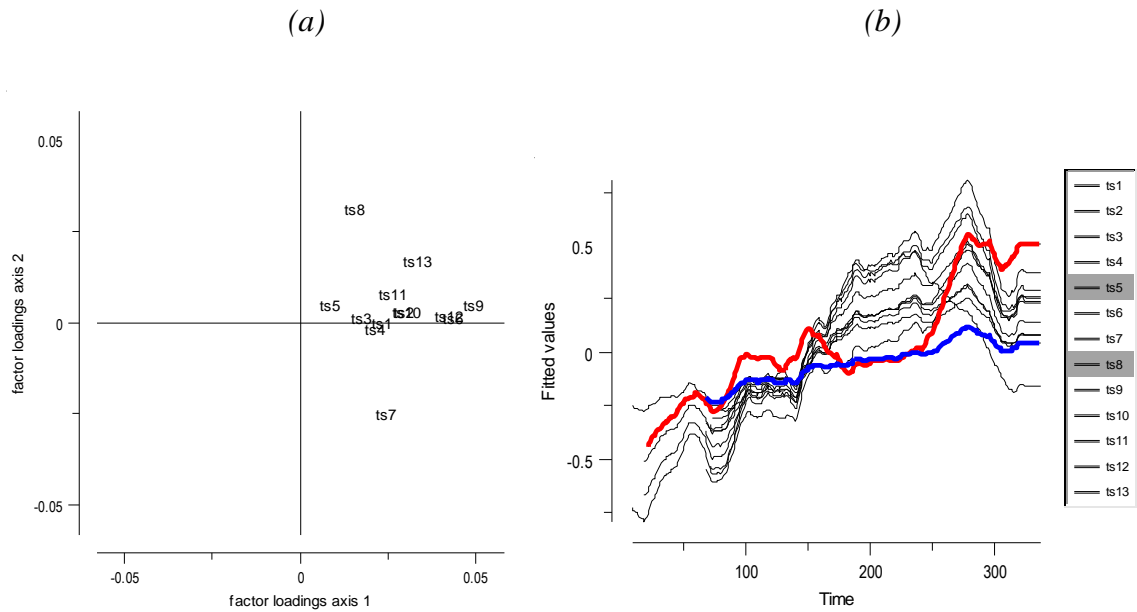


temperature and annual rainfall) is the most appropriate model to be fitted to the thirteen River Dee sites (highlighted in blue in Table 5.3.2.1).

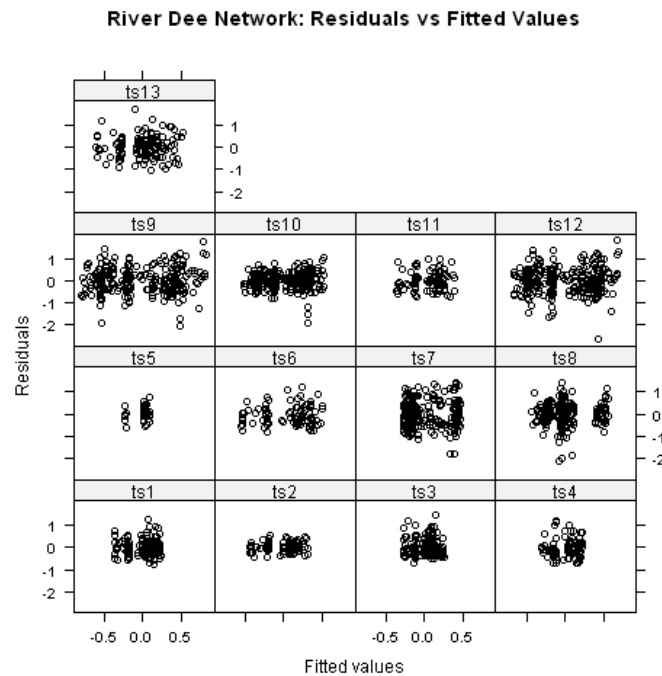
Figure 5.3.2.1 (a) displays the corresponding factor loadings – this plot suggests grouping of sites, pending on whether they lie closer to axis 1 or axis 2. The factor loadings suggest that sites 1, 2, 3, 4, 6, 7, 9, 10, 11, 12 are mainly driven by the first common trend; but, sites 5 and 8 are mainly driven by the second.

<b>Summary of Dynamic Factor Analysis Models Fitted to River Dee Network</b>				
<b>Diagonal Error Covariance Matrix</b>			<b>Non-Diagonal Error Covariance Matrix</b>	
No. Trends	Explanatory Variables	AIC	Explanatory Variables	AIC
1	-	2074.91	-	1765.41
2	-	2043.19	-	<b><u>1730.926</u></b>
3	-	2007.401	-	1760.776
1	Temperature	2079.257	Temperature	1765.364
1	Rain	2082.407	Rain	1765.365
1	Temperature and Rain	2087.27	Temperature and Rain	1765.21
2	Temperature	2056.836	Temperature	1731.058
2	Rain	2056.836	Rain	1731.012
2	Temperature and Rain	2047.404	Temperature and Rain	<b><u>1730.906</u></b>
3	Temperature	2008.899	Temperature	1752.529
3	Rain	2009.256	Rain	1750.962
3	Temperature and Rain	2008.288	Temperature and Rain	1762.348

**Table 5.3.2.1: Summary of Dynamic Factor Analysis models fitted to the 13 time series in the River Dee network.**



**Figure 5.3.2.1: Factor loadings corresponding to the two common trends (a); Fitted values obtained by the DFA model with two common trends. The blue line corresponds to site 5 and the red line corresponds to site 8 (sites 5 and 8 seem to influence one common trend); black lines correspond to the other sites which seem to influence the other common trend.**



**Figure 5.3.2.2: Residuals vs Fitted Values of the 13 time series from the final DFA model fitted.**

To demonstrate the differences between the groups of time series, the fitted values for each of the sites are displayed in Figure 5.3.2.1 (b). A blue and red line is used to represent sites 5 and 8, respectively in Figure 5.3.2.1 (b) – this highlights the coherence of their trends, but also, how their trends slightly differ from the other sites, especially after 150 months (approximately from the year 1995 onwards).

To check the validation of the final DFA model fitted to the River Dee sites, the residuals can be plotted against the fitted values for each of the thirteen time series, as seen in Figure 5.3.2.2. Zuur et al. (2007) state that as the  $n$  time series are being summarised by a small number of common trends, it is likely that validation plots, such as Figure 5.3.2.2 will show some patterns. Having said this, the residuals vs fitted values do not seem to show any strong trends or patterns.

## 5.4 River Dee Network Conclusion

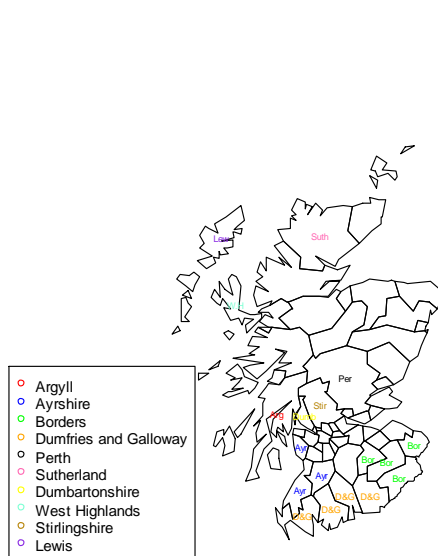
The Seasonal Mann Kendall test (Hirsch et al., 1982; Van Bell et al., 1984) and Dynamic Factor Analysis (Zuur et al., 2007) was applied to the thirteen River Dee network sites to measure the homogeneity of the sites. Based on the  $\chi^2$  value from the Seasonal Mann Kendall test, we failed to reject the null hypothesis that the trend at each site is in the same direction; but, the trend was only in the same direction during the seasons: winter, spring and autumn. The season seems to have a strong influence on the trend in the River Dee network. The use of Dynamic Factor Analysis was effective in providing a more detailed insight into the trends (but applied to the de-seasonalised data). The DFA highlighted that, based on AIC values, a DFA model fitted with two common trends, the inclusion of a non-diagonal error covariance matrix and the explanatory variables (annual mean temperature and annual rainfall) appropriately captured the coherency between the thirteen time series. Overall, the seasonal Mann-Kendall test leads one to believe that the trend of the log TOC at each of the

network sites is in the same direction; but, more specifically, the DFA suggests that there are actually two underlying common trends in the network. Furthermore, the DFA suggests that the explanatory variables annual mean temperature and annual rainfall possibly drive the observed trends.

## **5.5 Scottish Regions**

This section shall focus on investigating the trends of log TOC on a larger scale than previously explored. For rivers and lochs independently, the trends of log TOC shall be examined in a selection of different regions in Scotland. It is important to mention that the spatial groupings are not ecologically based, and that the specified regions are of different catchment and river basin sizes. SEPA has defined in which region of Scotland each river and loch site is located. The sites have been grouped based on SEPA's definition. The locations of the regions under scrutiny are displayed in Figure 5.5.1. With regards to river sites, the following regions shall be considered: Argyll, Ayrshire, Borders, Dumfries and Galloway, West Highlands, Perthshire and Sutherland. The regions concerning loch sites shall be: Dunbartonshire, West Highlands, Perthshire, Stirlingshire, Sutherland and Lewis. These particular regions were chosen for analysis, based on their reasonable number of sites situated within the region.

## Regions Investigated in Scotland



*Figure 5.5.1: Regions under investigation in Scotland.*

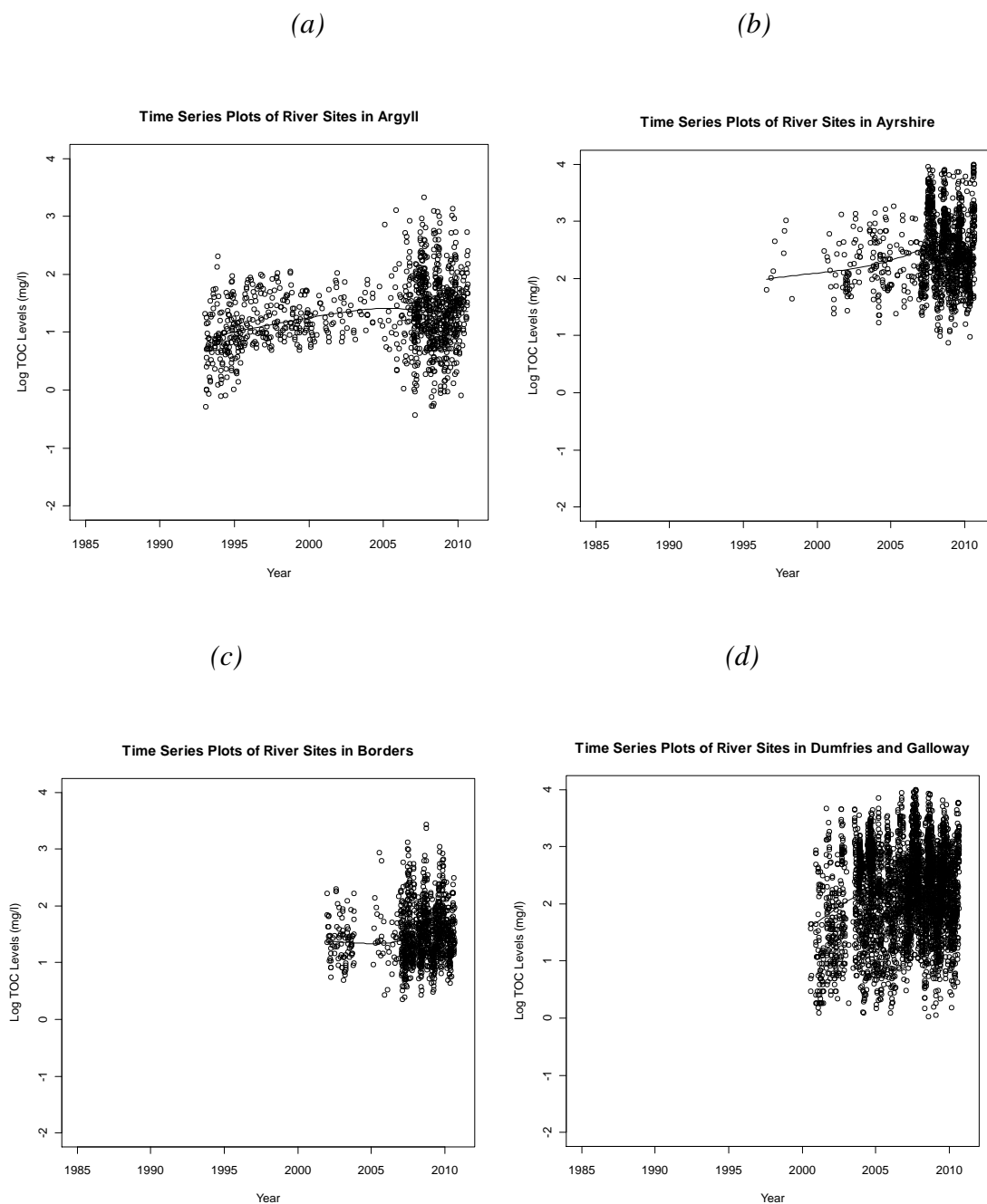
### 5.5.1 Initial Impression of Regions

The regions under investigation are summarized in Table 5.5.1.1. Time series plots are useful for gaining an initial impression of the coherency between sites situated in the same region. Based on the examination of Figures 5.5.1.1 to 5.5.1.3, it seems plausible (for both rivers and lochs) that sites situated in the same region, have log TOC trends which could be described as being coherent. The trends displayed, reinforce previous subjective impressions of rivers: the log TOC levels seem to increase up until the early 2000's, where the increase then either weakens or evens out. This trend seems to be stronger in the rivers sites, as expressed in earlier chapters. With regards to lochs, Figure 5.5.1.3 (a) suggests that log TOC levels are also increasing in the Dunbartonshire lochs from the late 1990's through until the mid-2000's; but, overall, Figure 5.5.3.1 highlights the unsteadiness of the log TOC levels in each region from 2005 onwards which was not emphasized in previous analysis.

Furthermore, it is evident from Figures 5.5.1.1 to 5.5.1.3 that the number of sites being monitored in the past five years has clearly increased – possibly explaining the increase in variability. The seasonality of log TOC within the regions was also considered. Figure 5.5.1.4 (a) displays the seasonality of the river sites log TOC levels in Argyll between 1994 and 2010; and to highlight the seasonal pattern in Argyll, Figure 5.5.1.4 (b) shows the seasonality in the year 2007. The exploratory analysis in the previous chapters highlighted the seasonal pattern of log TOC, which again is supported by Figures 5.5.1.4 (a) and (b). It is clear from Figure 5.5.1.4 that the variability increases from 2007 onwards – this is possibly due to an increase in number of sites being monitored in Argyll from this point in time. In Argyll, the levels of log TOC appear to increase from early spring until early autumn, which is then followed by a decrease - this is similar behaviour of rivers and lochs in the other regions.

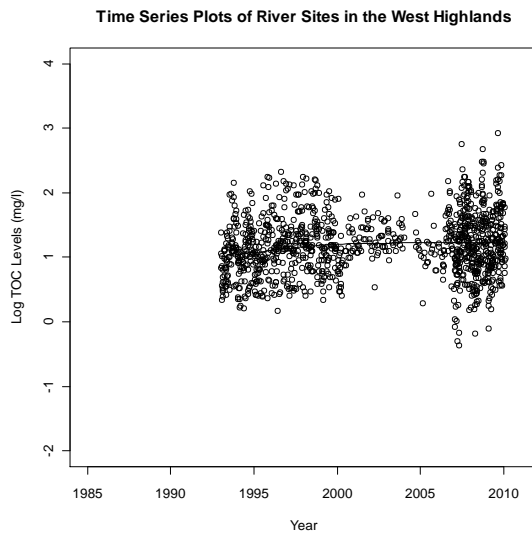
<b><u>Summary of Time Series in Regions</u></b>		
<b><u>Region</u></b>	<b><u>Number of Time Series</u></b>	<b><u>Longest Length of Time Series</u></b>
Rivers in Argyll	21	1993-2011
Rivers in Ayrshire	19	1997-2011
Rivers in Borders	23	2002-2011
Rivers in Dumfries & Galloway	57	2001-2011
Rivers in W. Highlands	13	1993-2011
Rivers in Perthshire	22	2007-2011
Rivers in Sutherland	17	1993-2011
Lochs in Dunbartonshire	8	1999-2011
Lochs in W. Highlands	7	2005-2011
Lochs in Perthshire	8	2005-2011
Lochs in Sutherland	9	2005-2011
Lochs in Lewis	16	2005-2011
Lochs in Stirlingshire	10	2006-2011

**Table 5.5.1.1: Summary of the river and lochs sites in each regio**

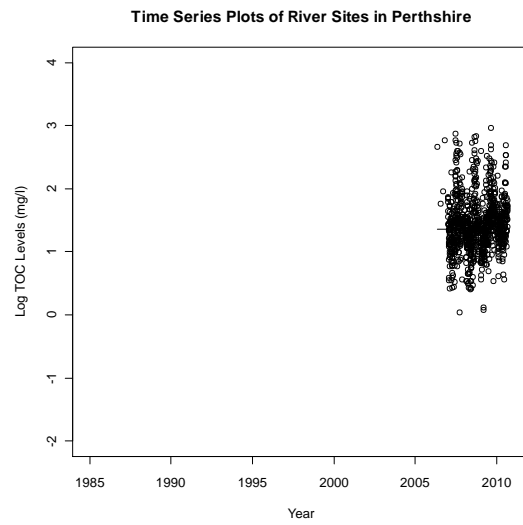


**Figure 5.5.1.1: Time series plots of log TOC in river sites at the Scottish regions: Argyll (a), Ayrshire (b), Borders (c) and Dumfries and Galloway (d).**

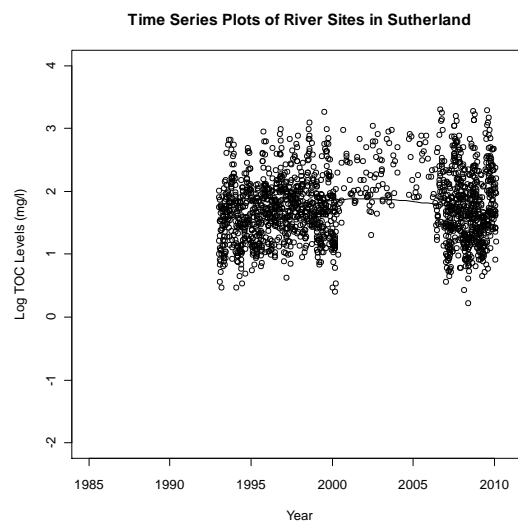
(a)



(b)

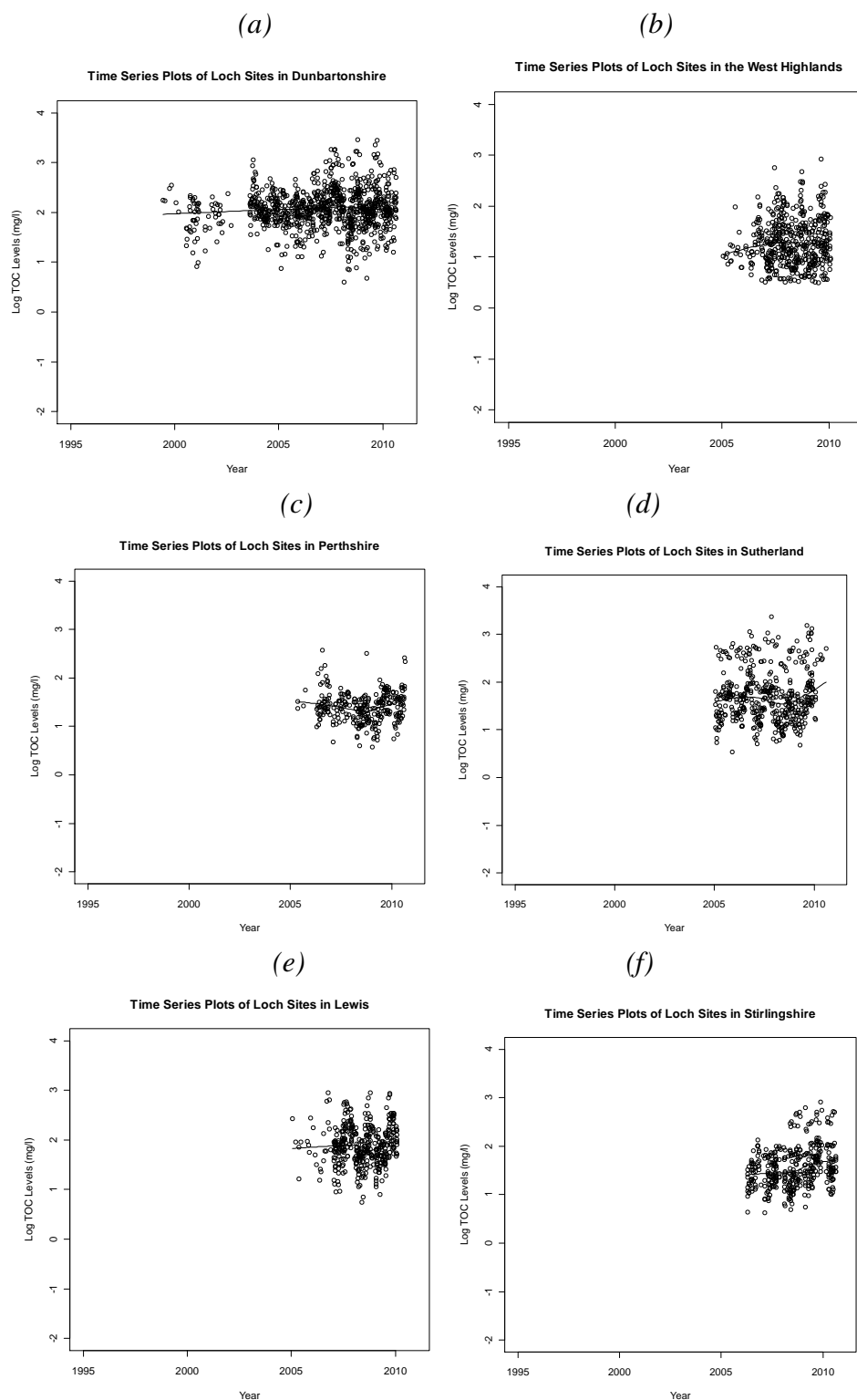


(c)



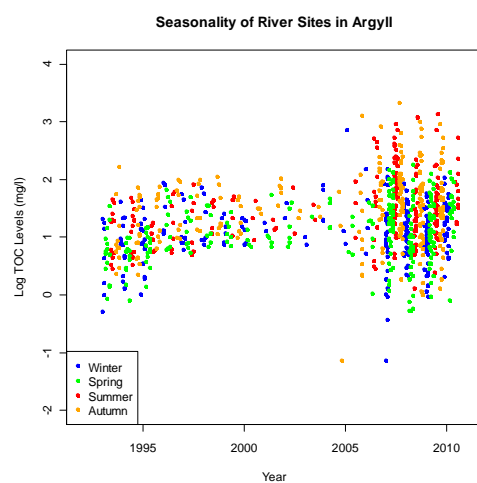
**Figure 5.5.1.2: Time series plot of log TOC in river sites in the Scottish regions: West Highlands (a), Perthshire (b) and Sutherland (c).**



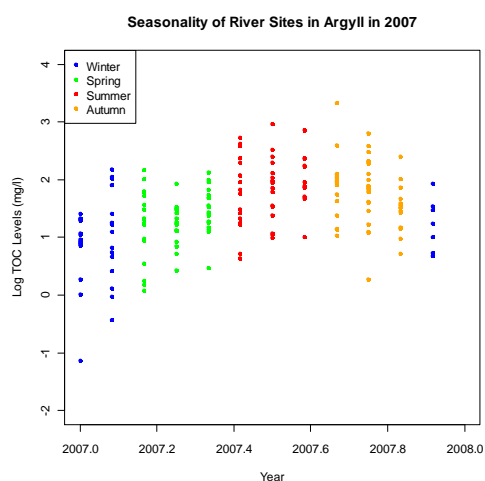


**Figure 5.5.1.3: Time series plot of log TOC in loch sites at the Scottish regions: Dunbartonshire (a), West Highlands (b), Perthshire (c), Sutherland (d), Lewis (e) and Stirlingshire (f)**

(a)



(b)



**Figure 5.5.1.4: Seasonality of log TOC levels in the Argyll rivers (a); and seasonality of log TOC levels in the Argyll rivers during the year 2007 (b).**

## 5.5.2 Applying the Seasonal Mann-Kendall Test to Scottish Regions

With regards to the Seasonal Mann Kendall test, it was decided, that it was only necessary to perform the test on a selection of regions, as this would give a further insight into the heterogeneity of sites located in the same region and a further understanding of the trends in each season. The rivers in the regions West Highlands and Perthshire; and the lochs in the regions Lewis and Sutherland were considered – these regions were assumed to be representatives of the other regions log TOC behaviour. Again, the methodology discussed in Section 5.2.1 was applied. The chi square test statistics and the corresponding chi square values from the Seasonal Mann Kendall analysis are summarised in Table 5.5.2.1.

The chi-square statistic in Table 5.2.1.1 ( $\chi^2_{station}$ ) can be applied to each region to test the null hypothesis that the trend at each site in the Scottish region is in the same direction. The chi-square statistic for each of the selected regions are displayed in Table 5.5.2.1. With regards to the heterogeneity of the stations, we fail to reject the null hypothesis for any of the specified regions, as the  $\chi^2_{station}$  statistics do not exceed the chi-square critical values [ $\alpha=0.05$ ] as Table 5.5.2.1 displays.

Similarly, the chi-square statistic in Table 5.2.1.1 ( $\chi^2_{season}$ ) can be applied to each of the specified regions to test the null hypothesis that the trend is in the same direction in each season. The  $\chi^2_{season}$  values for each region are displayed in Table 5.5.2.1 – again, the chi-square statistics were referred to the appropriate degrees of freedom, as outlined earlier in Table 5.2.2.1.

<b><u>Summary of Seasonal Mann Kendall Test</u></b>					
<b>Regions</b>	$\chi^2_{station}$	$\chi^2_{station}$ <b>Critical Value</b>	$\chi^2_{season}$	$\chi^2_{season}$ <b>Critical Value</b>	<b>Seasons which the trends are in significantly different directions</b>
West Highland Rivers	9.08	21.02	33.11	7.81	Winter, Summer and Autumn
Perthshire Rivers	14.42	32.67	55.48	7.81	Winter and Summer
Lewis Lochs	21.02	24.99	33.75	7.81	Winter and Spring
Sutherland Lochs	4.31	15.50	32.09	7.81	Winter and Summer

**Table 5.5.2.1: Summary of the Seasonal Mann Kendall tests performed on the specified regions.**

Since  $\chi^2_{season}$  was greater than the critical value of 7.81 in each region, the null hypothesis was rejected. Similar to the River Dee network, the  $\chi^2_{season}$  is significant, but,  $\chi^2_{station}$  is not, which means that the trends have significantly different directions in a different season or seasons, but not at different stations. Since this is the case, again, van Belle and Hughes (1984) chi-square statistic (5.3.1.1) can be applied to test the null hypothesis that there was a different trend direction in each season. Table 5.5.2.1 reveals that the direction of the trend is significantly different in the season ‘winter’ in each of the specified regions.

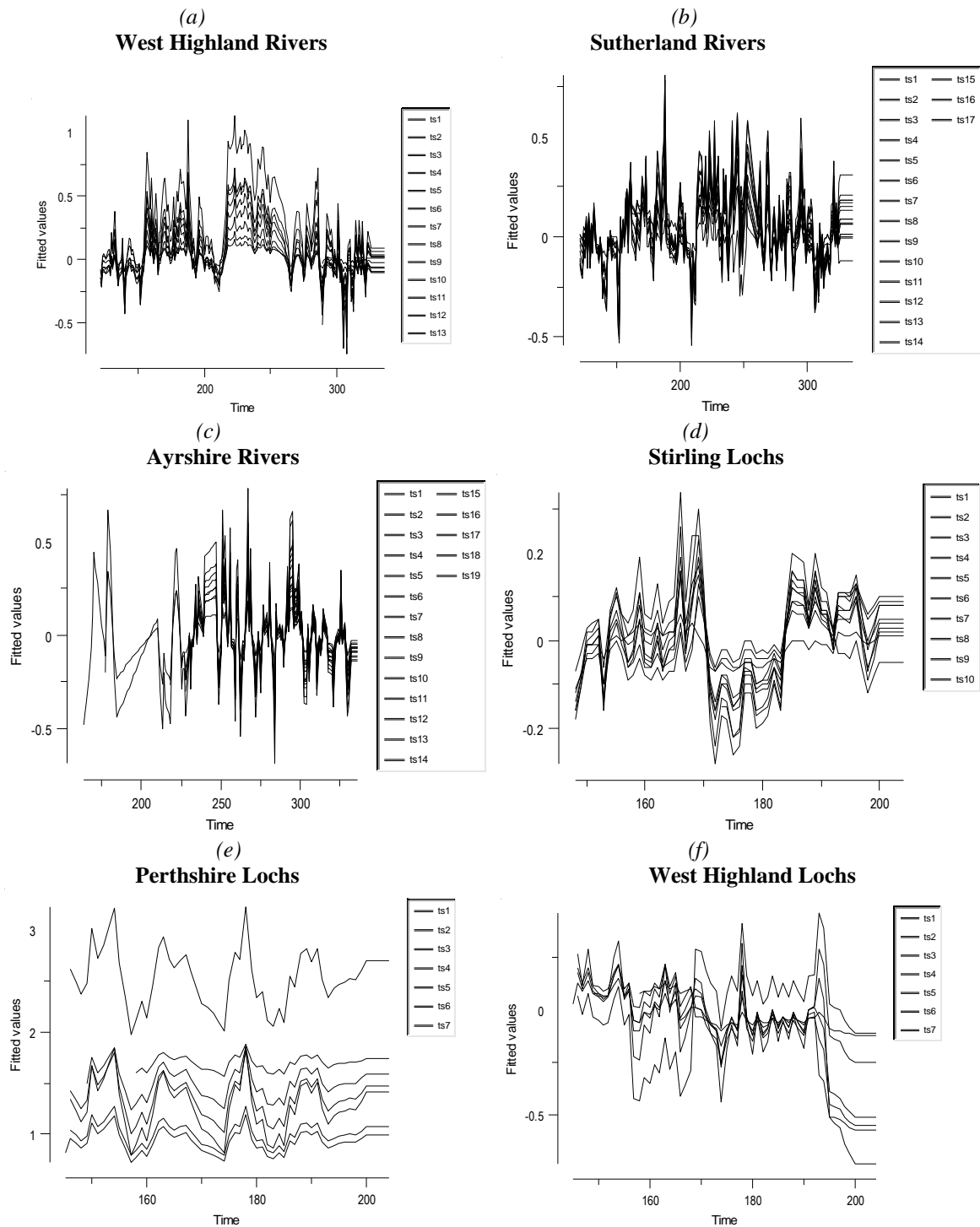
### 5.5.3 Applying the Dynamic Factor Analysis to the Scottish Regions

Similar to Section 5.3.2, DFA model were fitted to each of the regions which initially considered common trends using expression (5.2.2.4). Again, the DFA model were fitted to the  $n$  number of time series present in each region with a varying number of common trends and either incorporating a diagonal or non-diagonal error covariance matrix. Each time, the AIC was recorded. This procedure was conducted for the groups of rivers and lochs located in each of the specified regions. The DFA model with the lowest AIC value was chosen to be the “best” model. Note: unfortunately, the software *Brodgar 2.7.2* struggles to fit DFA models with more than 30 time series as using larger data sets means that the computing time becomes in the order of hours and the algorithm becomes unstable (Zuur, 2003a). Hence, a DFA model could not be fitted to the 57 river sites in Dumfries and Galloway. Table 5.5.3.1 summarises the number of common trends and error covariance matrix included in the final DFA models (which only considered common trends) fitted to each of the Scottish regions, for rivers and lochs, respectively.

Studying Table 5.5.3.1 highlights that all of the final DFA models fitted to the regions, only include one common trend. A DFA model fitted with one common trend, suggests that the log TOC levels in different sites (located in the same region) are behaving in a coherent fashion. Figure 5.5.3.1 displays a selection of plots – the fitted values in each of the plots were extracted from the final DFA models fitted to these particular regions (which included only one common trend). Figure 5.5.3.1 highlights the coherency between sites in each of the specified regions and supports the final DFA models only including one common trend.

<b><u>Summary of the DFA Models Fitted to Scottish Regions</u></b>			
<b><u>-Common Trends</u></b>			
<b><u>Rivers</u></b>	<b><u>Region</u></b>	<b><u>No. Trends</u></b>	<b><u>Error Covariance Matrix</u></b>
	Argyll	1	Diagonal
	Ayrshire	1	Non-Diagonal
	Borders	1	Non-Diagonal
	West Highlands	1	Diagonal
	Perthshire	1	Non-Diagonal
	Sutherland	1	Non-Diagonal
<b><u>Lochs</u></b>	Dunbartonshire	1	Non-Diagonal
	West Highlands	1	Non-Diagonal
	Perthshire	1	Diagonal
	Stirling	1	Non-Diagonal
	Sutherland	1	Non-Diagonal
	Lewis	1	Non-Diagonal

**Table 5.5.3.1: Summary of the final DFA models fitted to each of the Scottish Regions, for rivers and lochs.**



**Figure 5.5.3.1: Selection of plots with the fitted values obtained from the final DFA models with one common trend. River sites located in the regions West Highlands (a), Sutherland (b) and Ayrshire (c). Loch sites located in the regions Stirling (d), Perthshire (e) and West Highlands (f).**

Again, similar to the River Dee network, covariates were added to the added to the DFA models using expression (5.2.2.5) to try to explain what is driving the common trend at each region. DFA models were fitted to the regions with varying numbers of common trends, a diagonal or non-diagonal error covariance matrix and combinations of the explanatory variables mean temperature and annual rainfall. The AIC values of the DFA models only considering common trends were compared to the AIC values of those DFA models taking into account common trends and explanatory variables. The model with the lowest AIC value was taken to be the best DFA model for each region and are summarised in Table 5.5.3.2.

It is important to note, that nine out of the twelve regions studied, included either one or both of the explanatory variables in the final DFA models fitted. The DFA has been an effective method of measuring the coherency of log TOC levels between sites in these particular regions; but, it has also highlighted that temperature and/or rainfall could plausibly be driving the observed trends in a majority of the regions. In Section 1.2, possible factors driving trends were discussed – albeit, the focus was DOC. It is thought that an increase in temperature, leads to greater microbial activity and enhanced decomposition of peat and thus increased production of DOC (Worrall et al., 2004) – hence, it is possible that, in a similar manner, an increase in temperature could have a similar effect on TOC. Furthermore, Worrall et al., (2003) suggested that an increase in DOC could be possibly explained by a change in the flow path of rivers (as a result of heavy rainfall), allowing richer sources of DOC to be accessed – again, it is possible that heavy rainfall has a similar effect on TOC. These explanatory variables are not specific to each of the sites included in each of the regions, but their inclusion, indicates that environmental factors may be driving the observed trends, based on the DFA models.

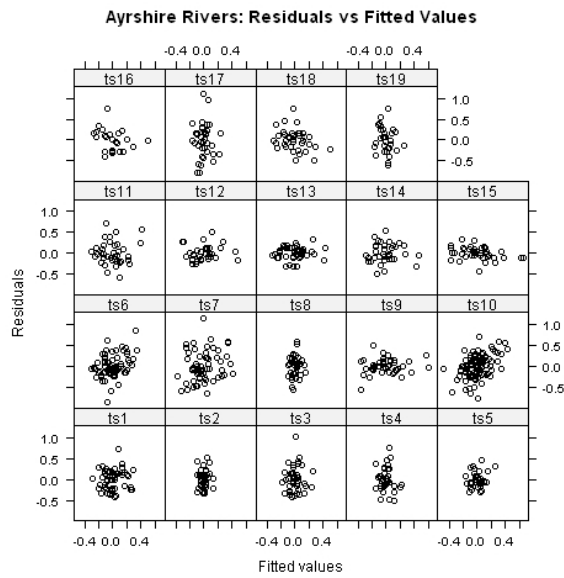
To check the validation of the final DFA models fitted to the regions, the residuals can be plotted against the fitted values – a selection are displayed in Figures 5.5.2.2 and 5.5.2.3. Having inspected the residuals vs fitted values plots displayed in Figures 5.5.2.2 and 5.5.2.3 and the residuals vs fitted values plots from the other DFA models, the plots do not seem to show any strong trends or patterns. Hence, the final DFA models seem to be appropriate.



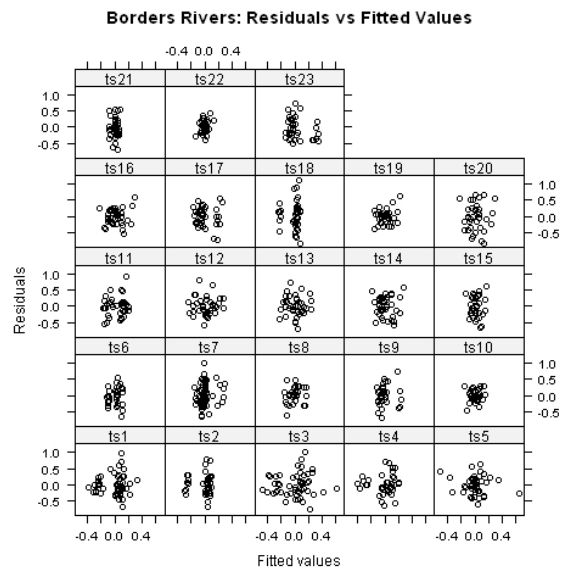
<b><u>Summary of the DFA Models Fitted to Scottish Regions</u></b> <b>- <u>Common Trends and Covariates</u></b>				
	<b><u>Region</u></b>	<b><u>No. Trends</u></b>	<b><u>Error Covariance Matrix</u></b>	<b><u>Explanatory Variables</u></b>
<b><u>Rivers</u></b>	Argyll	1	Diagonal	-
	Ayrshire	1	Non-Diagonal	Temperature and Rain
	Borders	1	Non-Diagonal	Temperature and Rain
	West Highlands	1	Diagonal	-
	Perthshire	1	Non-Diagonal	Temperature and Rain
	Sutherland	1	Non-Diagonal	Rain
<b><u>Lochs</u></b>	Dunbartonshire	1	Non-Diagonal	-
	West Highlands	1	Non-Diagonal	Temperature
	Perthshire	1	Diagonal	Rain
	Stirling	1	Non-Diagonal	Temperature
	Sutherland	1	Non-Diagonal	Temperature and Rain
	Lewis	1	Non-Diagonal	Temperature and Rain

**Table 5.5.3.2: Summary of the final DFA models fitted to each of the Scottish Regions, for rivers and lochs.**

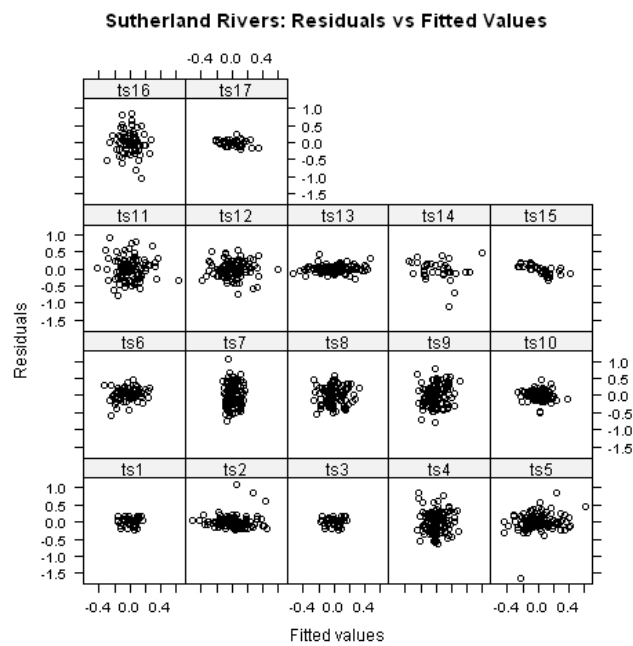
(a)



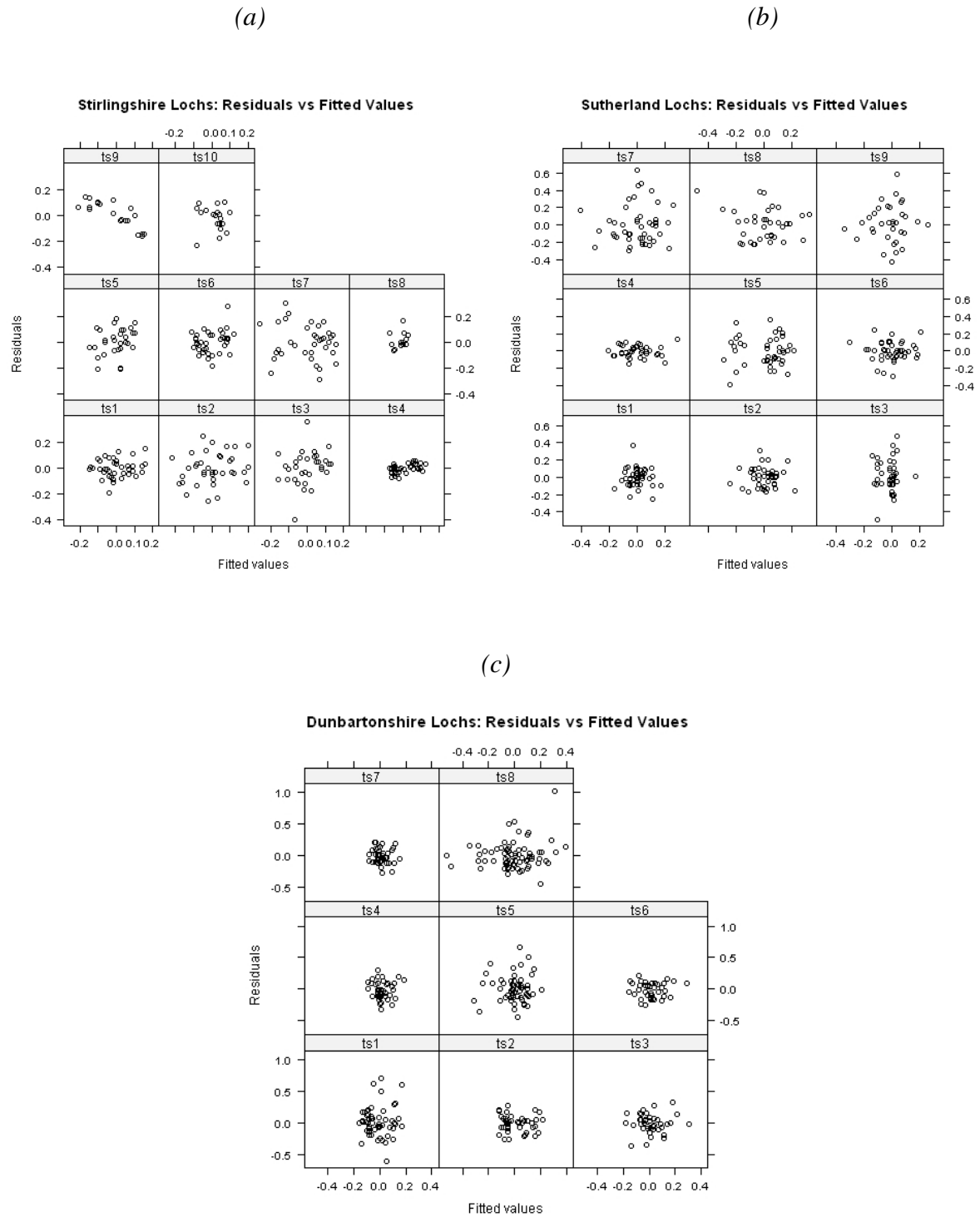
(b)



(c)



**Figure 5.5.3.2: A selection of residuals vs fitted values plots from the final DFA models fitted to the river sites in the regions: Ayrshire (a), Borders (b) and Sutherland (c).**



**Figure 5.5.3.3: A selection of residuals vs fitted values plots from the final DFA models fitted to the loch sites in the regions: Stirlingshire (a), Sutherland (b) and Dunbartonshire (c).**

## 5.6 Conclusion

The literature review in this chapter highlighted the variety of ways in which the problem of coherency can be tackled. Having studied the literature, it was thought, that dynamic factor analysis and the seasonal Mann-Kendall test were appropriate techniques to be applied in this thesis.

Having considered the River Dee network in great detail in the previous chapter and finding the sites to be spatially independent (Section 4.4.2), it was of interest to measure the coherency between the log TOC levels at the thirteen sites. The seasonal Mann-Kendall test and dynamic factor analysis were applied to the thirteen sites. Based on the Seasonal Mann Kendall test, it was concluded that the trend at each of the thirteen sites was in the same direction; but, the trend was not in the same direction in each of the seasons. It was concluded that the trend was in the same direction in the seasons: winter, spring and autumn; but, the trend was not in the same direction during the summer season. Dynamic factor analysis models were then fitted to the thirteen sites with varying number of common trends, the inclusion of an error covariance matrix (diagonal or non-diagonal) and the inclusion of explanatory variables common to all sites (annual mean temperature and annual rainfall). The AIC values of each DFA model were compared and the final DFA model fitted to the thirteen River Dee sites included two common trends, the inclusion of a non-diagonal error covariance matrix and the explanatory variables annual mean temperature and annual rainfall (for the data relevant to the East of Scotland). Interpreting the results from the analysis would lead one to believe that overall, the log TOC at each of the network sites is behaving coherently; but, more specifically, there are actually two underlying common trends in the network. Furthermore, it seems possible that the annual mean temperature and annual rainfall in the east of Scotland were driving the increase in log TOC in the River Dee network between 1990 and 2006 (Worral et al., 2003; Worral et al., 2004).

This chapter then moved on to considering rivers and lochs on a larger scale than the analysis conducted previously in the thesis. Regions of Scotland were investigated.

Based on exploratory analysis, it seemed plausible (for both rivers and lochs) that sites situated in the same region, have log TOC trends which could be described as being coherent. The trends displayed, supported previous subjective impressions of rivers: the log TOC levels seemed to increase up until the early 2000's, where the increase then either weakened or flattened out. This trend seems to be stronger in the rivers sites, as expressed in earlier chapters. However, exploring the trends of the lochs in different regions suggested that log TOC levels in Dunbartonshire behave similarly to the rivers in regions; but, overall, from 2005 onwards the log TOC levels become fairly unsteady in each region. The seasonality of log TOC within the regions was also considered and was found to mirror the seasonal patterns seen previously.

Similar to the River Dee sites, a seasonal Mann-Kendall test was applied to a selection of the regions and DFA was performed to gain an understanding of the coherency of log TOC levels in different sites located in the same region. The seasonal Mann-Kendall test was performed on the rivers in the regions West Highlands and Perthshire; and the lochs in the regions Lewis and Sutherland were considered. For each of these regions, it could be concluded that the trend of the sites was in the same direction; but, similar to the River Dee sites, the trend was not in the same direction in each of the seasons, suggesting that the season could also be a strong driver of trend in the regions.

Dynamic factor analysis models were then fitted to each of the regions – again, with varying number of common trends, the inclusion of an error covariance matrix (diagonal or non-diagonal) and the inclusion of explanatory variables common to all sites (annual mean temperature and annual rainfall). All of the final DFA models fitted included one common trend. This suggests that the log TOC levels of river and loch sites located in the same region, behave coherently. Also, nine out of the twelve regions studied, included either one or both of the explanatory variables in the final DFA models fitted. Even though the explanatory variables included in the DFA were not site specific, their inclusion in nine out of

the twelve final DFA models, suggests that environmental factors such temperature and rainfall, appear to influence the trends of log TOC in the majority of regions.

Having explored the coherency between sites in each region, the next chapter focuses on appropriately modelling the log TOC levels in each region – taking into consideration the trend and seasonality, over time and space.

# **Chapter 6**

## **Modelling Log TOC, Over Time and Space in Scottish Regions**

The previous chapter explored the coherency between sites located in the same region. Taking into consideration the results from the coherency analysis, this chapter aims to build a model which appropriately captures the behaviour of log TOC for the rivers and lochs in each of the regions (specified in the previous chapter). This chapter shall fit additive models to each of the regions to capture the trend and seasonality of the log TOC, over time and space. The inclusion of covariates in the additive models shall be explored, in an attempt to explain the observed trends and patterns in each of the regions.

## 6.1 Modelling Trend and Seasonality

Based on the visual inspection of the time series plots in Figures 5.5.1.1-5.5.1.3 displayed in Chapter 5, non-parametric regression seems to be the most appropriate method of capturing the behaviour of log TOC in each of the regions, for both rivers and lochs. Hence, similar to previous chapters, a GAM model shall be fitted to each of the regions.

Initially, the trend and seasonality of the regions shall be considered by fitting additive models which take into consideration time and space. The ‘Site’ shall be included in the GAM models fitted to capture the ‘space’ element across the region. The time and space shall be incorporated in the GAM models by fitting an interaction between year and site, but also, month and site. The spatial coordinates were used in Chapter 4 for the analysis of the River Dee network; however, including the ‘site’ essentially serves the same purpose in the regions. Due to the previous analysis, the presence of a seasonal pattern is evident, and therefore shall be incorporated in the models fitted. Therefore, letting  $y = \log \text{ TOC level at a site}$ ;  $\text{Year} = \text{Year}$ ;  $\text{Month} = \text{Month}$ ; and each Site in the region = Site, the following additive model can be fitted, which incorporates trend, seasonality and the time and space interactions (still assuming the  $\varepsilon_i$  are independent with mean 0 and constant variance  $\sigma^2$ ) :

$$y_i = \beta_0 + m_1(\text{Year}_i) + m_2(\text{Month}_i) + m_3(\text{Site}_i) + m_4(\text{Year}_i * \text{Site}_i) + m_5(\text{Month}_i * \text{Site}_i) + \varepsilon_i$$

$$i = 1, \dots, n \quad (6.1.1)$$

Again, the degree of smoothing applied to each term was chosen by generalized cross validation. In expression (6.1.1), the  $\varepsilon_i$  are still assumed to be independent based on the spatial dependence analysis performed in Section 4.4.2 – the River Dee sites in the same network were deemed to be spatially independent; therefore, it seems plausible that sites located in a larger geographical space, will also be spatially independent.

Tables 6.1.1 and 6.1.2 summarise the final trend and seasonality GAM models fitted to each of the regions, with regards to rivers and lochs. Note: a GAM model was also fitted to the 57 rivers in the region Dumfries and Galloway – unlike the DFA, a large number of time series did not present a problem. Studying Tables 6.1.1 and 6.1.2 highlights that the trend and



seasonality terms fitted in each GAM model (for each region independently) are significant with p-values less than 0.05. The exceptions being, the lochs in Stirlingshire (Year term not significant) and the lochs in Sutherland (Year and Month term not significant); however, if the interaction term including that particular term was significant, standard practise is for the term to remain in the model.

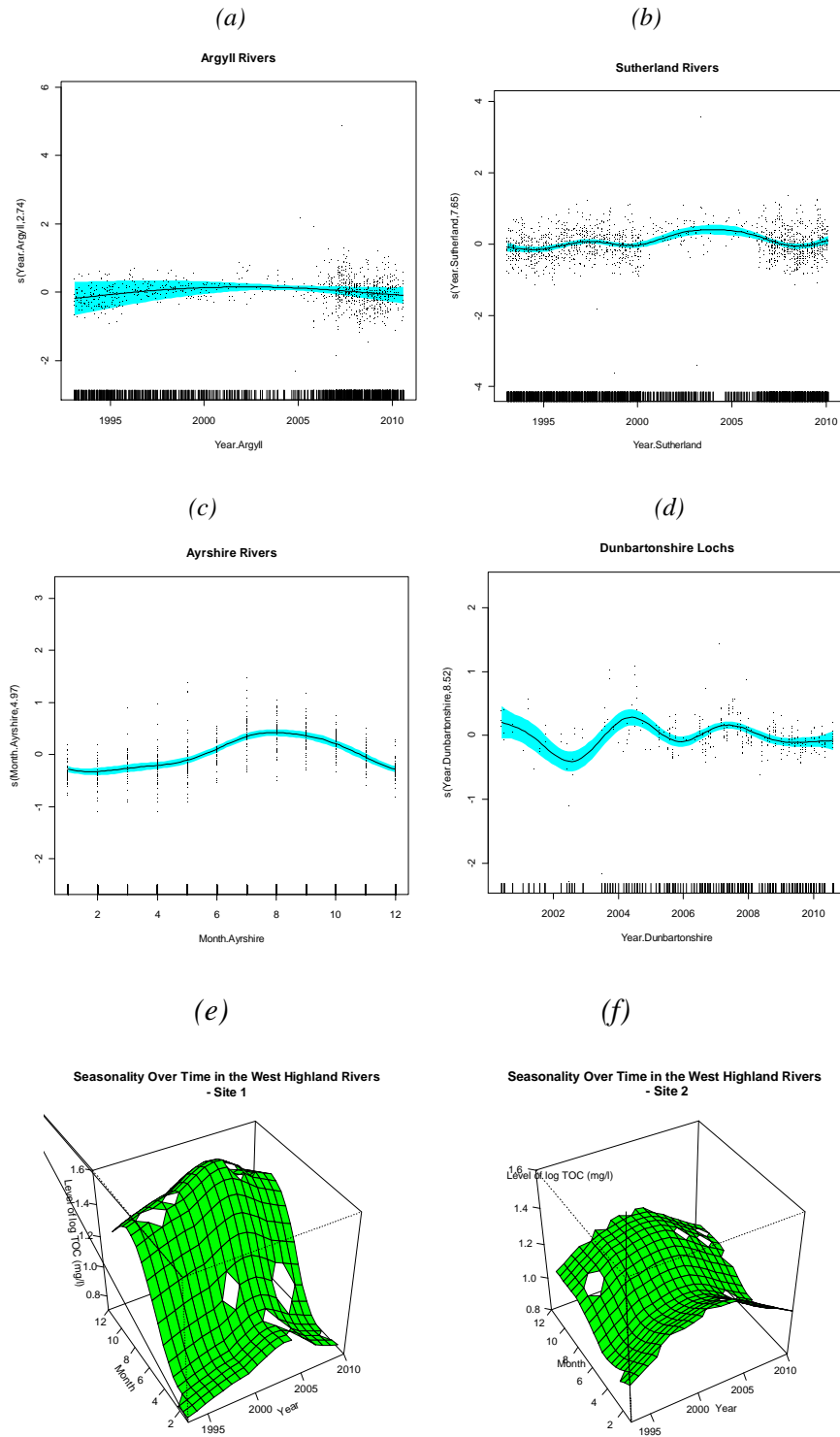
Figure 6.1.1 displays a selection of the effect plots from the fitted trend and seasonality GAM models. Figure 6.1.1 (a) and (b) display the trend at the rivers located in Argyll and Sutherland, respectively. Similar to the previous initial impressions formed, it appears, that as the log TOC levels increase between the early 1990's and early 2000's – after this time period, the log TOC levels seem to level out. This trend is similar in the rivers located in the regions: the Borders, Dumfries and Galloway, West Highlands and Perthshire. Figure 6.2.3.2 (d) displays the trend of log TOC in the Dunbartonshire Lochs. The longest time series available for the lochs is in Dunbartonshire – the log TOC levels appear to decrease in the late 1990's, before sharply increasing until the mid-2000's. After the year 2005, the log TOC levels appear to be fairly unsteady – showing signs of increasing and decreasing over the remaining years. This trend from 2005 onwards was similar in lochs located in the West Highlands, Perthshire, Sutherland, Stirlingshire and Lewis. Figure 6.1.1 (c) displays the seasonality of log TOC in the Ayrshire rivers – a pattern which is similar in all the rivers and lochs located in the other regions.

<b>Summary of GAM Models Fitted to River Sites in Each Region</b>						
	<b>Pr(&gt; t )</b>					
Region	Year	Month	Site	Year*Site	Month*Site	Adjusted R.Sq
Argyll	<0.001	<0.001	<0.001	<0.001	<0.001	47%
Ayrshire	<0.001	<0.001	<0.001	<0.001	<0.001	51.6%
Borders	<0.001	<0.001	<0.001	<0.001	<0.001	29.6%
Dumfries and Galloway	<0.001	<0.001	<0.001	<0.001	<0.001	33.7%
West Highlands	<0.001	<0.001	<0.001	<0.001	<0.001	57%
Perthshire	<0.001	<0.001	<0.001	<0.001	<0.001	30.3%
Sutherland	<0.001	<0.001	<0.001	<0.001	<0.001	40.7%

**Table 6.1.1: Summary of the trend and seasonality GAM models fitted to the river sites located in the different regions of Scotland.**

<b>Summary of GAM Models Fitted to Loch Sites in Each Region</b>						
	<b>Pr(&gt; t )</b>					
Region	Year	Month	Site	Year*Site	Month*Site	Adjusted R.Sq
Dumbartonshire	<0.001	<0.001	<0.001	<0.001	<0.001	27.2%
West Highlands	0.03	<0.001	<0.001	<0.001	<0.001	82.5%
Perthshire	<0.001	<0.001	<0.001	<0.001	<0.001	62.9%
Sutherland	0.15(P)	0.32	<0.001	<0.001	<0.001	61%
Stirlingshire	0.06	0.01	<0.001	<0.001	0.009	60%
Lewis	0.41	<0.001	<0.001	<0.001	<0.001	42%

**Table 6.1.2: Summary of the trend and seasonality GAM models fitted to the loch sites located in the different regions of Scotland. Note: all the terms in the final model are included as non-parametric terms, except from those marked with a “(P)” – these terms are parametric.**



**Figure 6.1.1: A selection of effects plots from the fitted trend and seasonality GAM models: year at the Argyle rivers (a), year at the Sutherland rivers (b), month at the Ayrshire rivers (c), year at the Dunbartonshire Lochs (d). Trend and seasonality 3D plots in the West Highland rivers site 1 (e) and site 2 (f).**

Tables 6.1.1 and 6.1.2 highlight that the ‘Site’ term and interaction terms included in the GAM models are significant in each of the regions. In a similar fashion to Figure 4.4.7.1 (Chapter 4), 3D trend and seasonality plots have been used to support the inclusion of the significant interaction terms in the GAM models. Two river sites from the West Highland region have been selected and are shown in Figure 6.1.1 [(e) and (f)]. Comparing ‘site 1’ to ‘site 2’ shows the seasonal pattern of the log TOC at both sites; but, highlights that the increase in log TOC levels between spring and autumn is more rapid in ‘site 1’ than ‘site 2’. With regards to trends, log TOC levels appear to increase up until the year 2005 in ‘site 1’ before decreasing; log TOC levels seem to smoothly increase up until the year 2003, before decreasing in ‘site 2’. These slight differences in trend and seasonal patterns could be observed between sites in each of the regions. These significant interaction terms imply that even if the log TOC levels of sites in a particular region are behaving coherently, it is still plausible that the levels differ between sites throughout the seasons and over the years.

Table 6.1.2 highlights that the GAM models fitted to the loch sites in the West Highlands is a very good fit to the data – with an adjusted R squared value of 82.5%. However, having fitted a GAM model which considers time and space, it is of interest to incorporate covariates, to see if they improve the GAM models fitted to each region.

## 6.2 Modelling Trend, Seasonality and Covariates

Section 6.1 identified that the terms: Year, Month, Site, Year\*Site and Month\*Site were included in the models fitted to each of the regions. A natural progression from this is to build a model, which captures the trend and seasonality of log TOC in each of the regions, over time and space, but also, incorporate covariates. Hence, the GAM models will be re-fitted to the regions, but this time, include the covariates temperature, pH, log alkalinity, log nitrate, log sulphate and the annual rainfall (mm) as it seems to be a sensible progression from Chapter 5. Each covariate added to the model shall be site specific, except from the annual rainfall unlike the mean annual temperature and annual rainfall included in the DFA models fitted in Chapter 5.

Letting  $y = \log \text{ TOC level at a site}$ ;  $\text{Year} = \text{Year}$ ;  $\text{Month} = \text{Month}$ ; each site in the region =  $\text{Site}$ ,  $T = \text{temperature}$ ;  $A = \log \text{ alkalinity}$ ;  $\text{pH} = \text{pH}$ ;  $S = \log \text{ sulphate}$ ;  $N = \log \text{ nitrate}$ ;  $R = \text{Annual Rainfall (mm)}$ , the following additive model can be fitted, (again, assuming the  $\varepsilon_i$  are independent with mean 0 and constant variance  $\sigma^2$ ) :

$$y_i = \beta_0 + m_1(\text{Year}_i) + m_2(\text{Month}_i) + m_3(\text{Site}_i) + m_4(T_i) + m_5(A_i) + m_6(\text{pH}_i) + m_7(S_i) + m_8(N_i) + m_9(R_i) + m_{10}(R_i) + m_{11}(\text{Year}_i * \text{Site}_i) + m_{12}(\text{Month}_i * \text{Site}_i) + \varepsilon_i$$

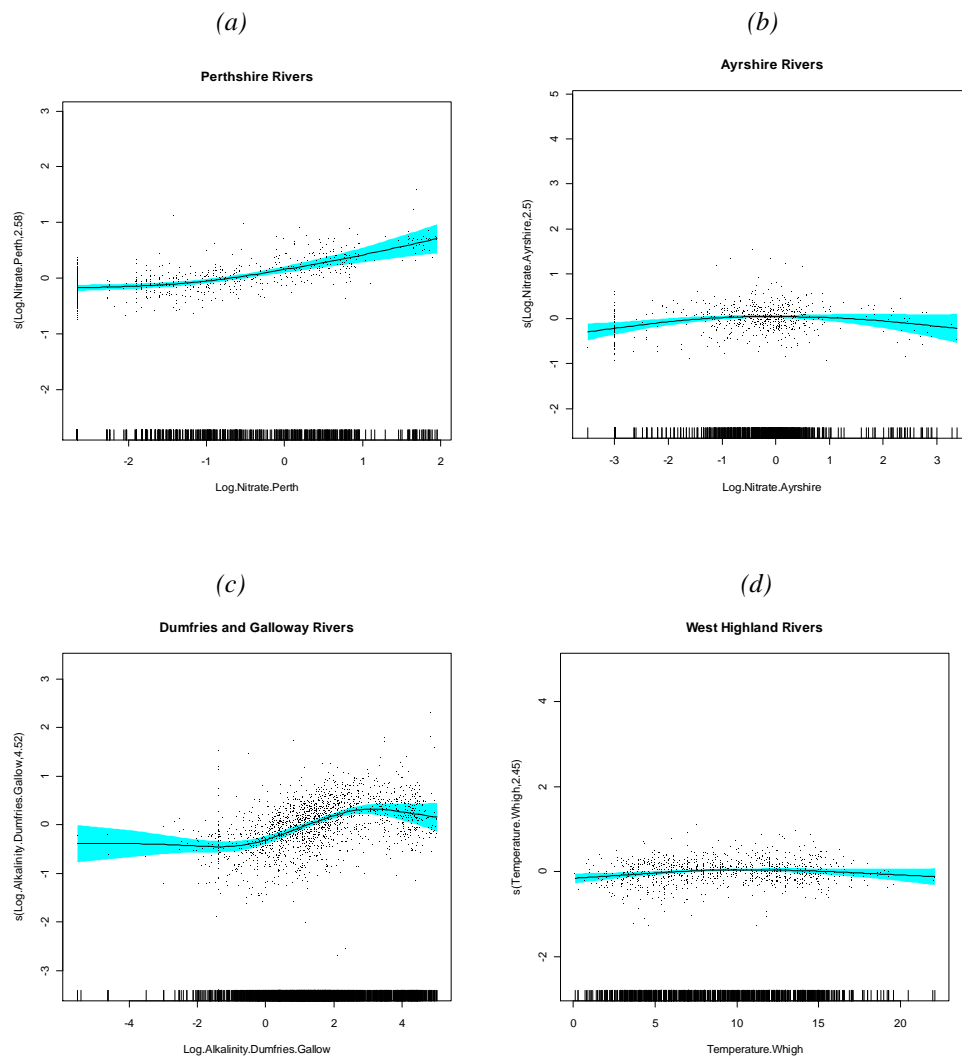
$$i = 1, \dots, n \quad (6.2.1)$$

Terms that were not significant at the 5% level were removed from the GAM model, and the model was refitted. A summary of the final GAM models fitted to each of the regions is presented in Tables 6.2.1 and 6.2.2. Including the covariates has not altered the inclusion of the terms ‘Year’, ‘Month’, ‘Site’ or the interaction terms in the GAM models fitted – these terms remain in the final GAM models fitted for both rivers and lochs in each of the regions.

The rivers shall be considered first. Based on the adjusted R-squared values, Table 6.2.1 reveals that including covariates in the GAM models fitted to the regions, has improved the models – with the exception being the rivers in Argyll. When expression (6.2.1) was fitted to the rivers in Argyll, the only covariate which was significant in the model was log nitrate; however, the GAM model including log nitrate had an adjusted R-squared value of 34.7%, compared to an adjusted R-squared value of 47% when expression (6.2.1) was fitted without any covariates (as seen in Table 6.2.1). This contrast in adjusted R-squared values was due to the missing log nitrate values – GAM model are fitted using ‘complete row analysis’, hence missing covariate values will have an effect on the fitted model and the amount of variation in the data which the model explains.

Summary of the Final GAM Models Fitted to Rivers in Regions													
Rivers	Region	Pr(> t )											Adjusted R. Sq
		Year	Month	Site	Year*Site	Month*Site	Temperature	pH	Log Alkalinity	Log Nitrate	Log Sulphate	Annual Rainfall	
	Argyll	<0.001	<0.001	<0.001	<0.001	<0.001	-	-	-	-	-	-	47%
	Ayrshire	<0.001	<0.001	<0.001	<0.001	<0.001	-	-	<0.001	0.004	<0.002 (P)	<0.001	63.1%
	Borders	<0.001	<0.001	<0.001	0.04	<0.001	<0.001	<0.001	<0.001		<0.001	-	46.2%
	Dumfries and Galloway	<0.001	<0.001	<0.001	<0.001	<0.001	-	<0.001	<0.001	<0.001	<0.001	-	56.2%
	West Highlands	0.002	<0.001	<0.001	<0.001	<0.001	0.005	-	<0.001	-	-	0.006	62%
	Perthshire	<0.001	<0.001	<0.001	<0.001	<0.001	0.002	-	<0.001	<0.001	<0.001	-	73.5%
	Sutherland	<0.001	<0.001	<0.001	<0.001	<0.001	-	0.02 (P)	-	-	0.01 (P)	-	53%

**Table 6.2.1: Summary of the final GAM models fitted to rivers in the specified regions. Note: if a term was not included in the final GAM model, it is represented by “-”; all the terms in the final model are included as non-parametric terms, except from those marked with a “(P)” – these terms are parametric.**



**Figure 6.2.1: A selection of effect plots from the final GAM models fitted- log nitrate in the Borders (a); log nitrate in Ayrshire (b); log alkalinity in Dumfries and Galloway (c); temperature in the West Highlands (d).**

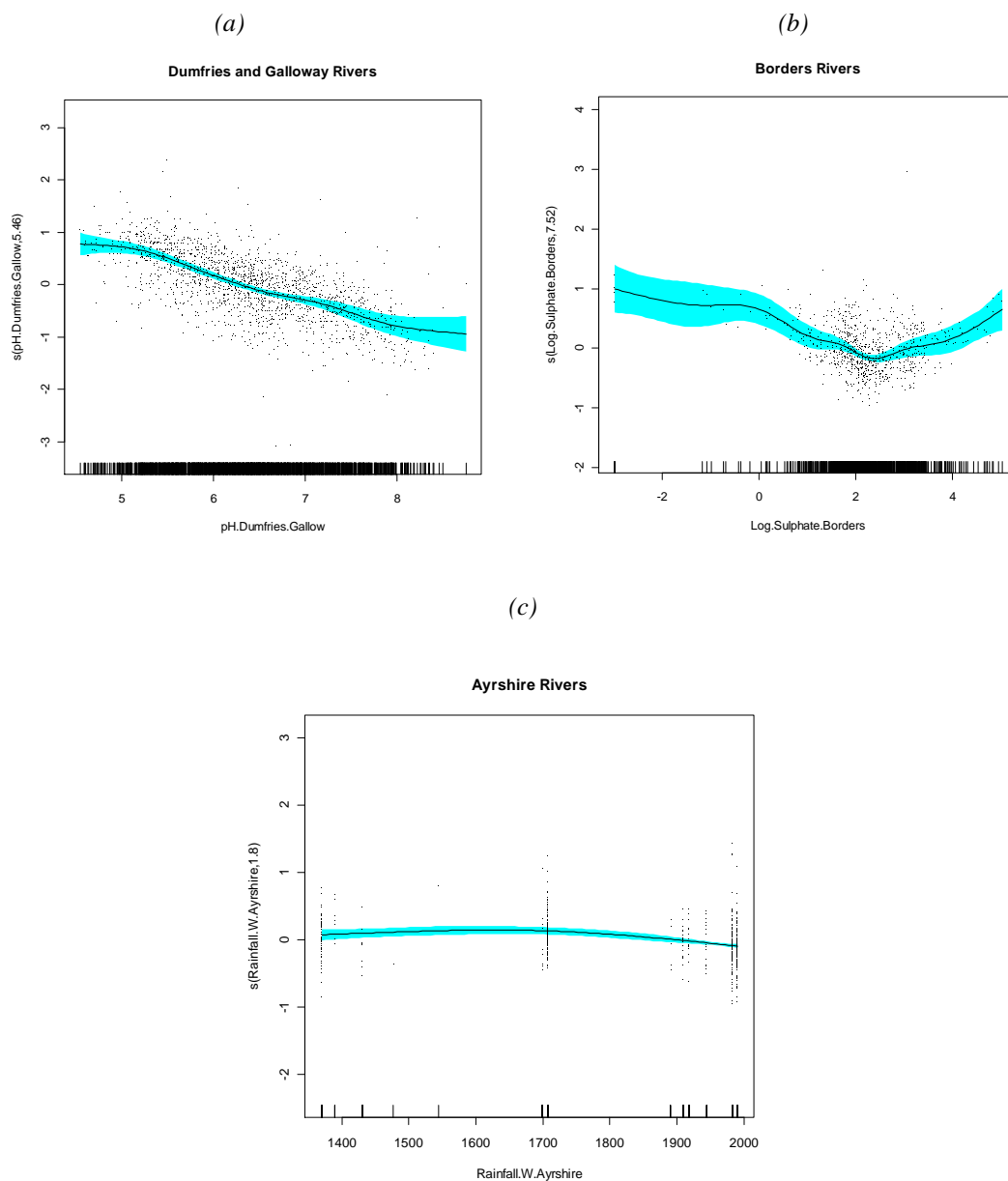
Studying Table 6.2.1 reveals that none of the covariates fitted in the GAM models are significant in all of the specified regions. However, log alkalinity and log sulphate are significant in 5 out of the 7 final GAM models fitted; pH, log nitrate and temperature are significant in 3 out of the 7 regions; and annual rainfall is significant in 2 out of the 7 regions. Only 7 of the Scottish regions have been considered here; however, the final GAM models fitted to the rivers suggests that the covariates considered here may not be able to explain the trends and patterns of log TOC in all seven regions, but, may be responsible for what is driving the trends and patterns in some of the regions. Figures 6.2.1 and 6.2.2 display a selection of the effect plots from the fitted GAM models – particularly focussing on covariates which seem to have the greatest effect on log TOC levels across the regions, based on the final GAM models fitted (as seen in Table 6.2.1).

Increasing log nitrate levels appear to have a different effect on log TOC levels in each of the regions, with regards to rivers. For example, Figure 6.2.1 [(a) and (b)] shows a contrasting effect of increasing log nitrate levels in the regions Perthshire, and Ayrshire. An increase in log nitrate levels in the Perthshire rivers seems to be associated with a smooth increase in the log TOC levels; but, an increase in log nitrate levels in the Dumfries and Galloway rivers, is associated with a smooth decrease in log TOC levels. An increase in the Ayrshire rivers is associated with an initial increase in log TOC levels followed by a gentle decrease.

The effect of increasing log alkalinity levels in Dumfries and Galloway, displayed in Figure 6.2.1 (c), is similar in Ayrshire, the Borders, West Highlands and Perthshire (with regards to rivers): the log TOC levels appear to remain fairly steady or faintly increase when log alkalinity levels increase up to (approximately) 2.5; however, when log alkalinity levels exceed (approximately) 2.5, log TOC levels seems to decrease smoothly.

Similar to the initial impressions gained in earlier chapters, Figure 6.2.1 (d) displays the effect of increasing temperature levels in the rivers located in the West Highlands. As temperature increases to approximately 12 degrees Celsius, log TOC also appears to increase; if temperature levels rise above approximately 12 degrees Celsius, log TOC levels appear to fall. This pattern is similar in the Borders and Perthshire.





**Figure 6.2.2: A selection of effect plots from the final GAM models fitted- pH in Dumfries and Galloway (a); log sulphate in the Borders (b); and annual rainfall in Ayrshire (c).**

Figure 6.2.2 (a) displays the effects of increasing pH levels in the rivers located in Dumfries and Galloway. It seems that as the pH levels increase, the levels of log TOC appear to decrease smoothly (a similar effect can be observed in the Borders) - this could be because pH is higher at sites with less peaty soils (less peaty = lower carbon content). The rivers in Sutherland and Perthshire behave in a similar manner; however, the decrease in log TOC levels appears to be linear.

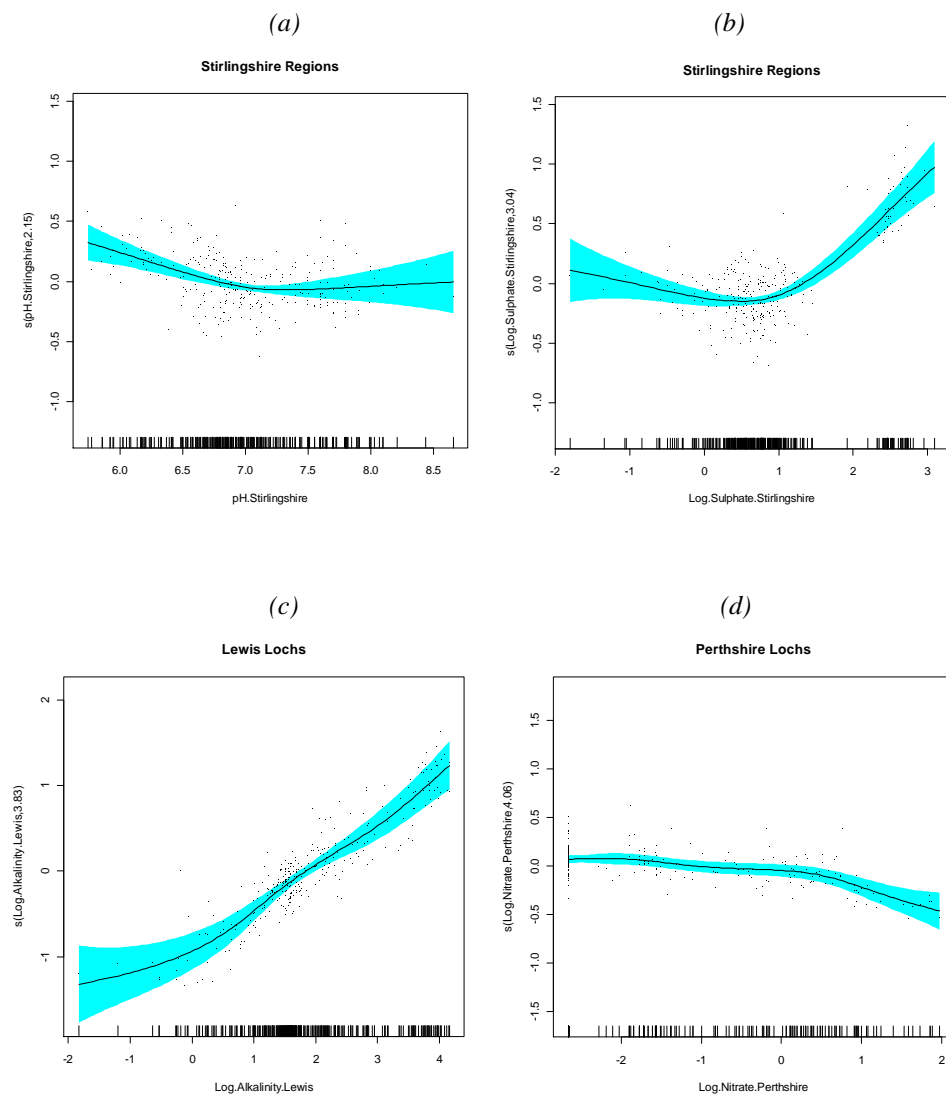
Figure 6.2.2 (b) displays the effects of increasing log sulphate levels in the rivers located in the Borders. As log sulphate levels increase up to approximately the value of 2 mg/l, log TOC levels seem to gradually decrease; however, if the log sulphates increase above approximately 2 mg/l, the log TOC levels seem to gradually increase – this pattern is similar in Perthshire and Dumfries and Galloway. This is not the case in Sutherland though, increasing log sulphate levels seem to be associated with a linear decrease in log TOC levels; and it is the opposite in Ayrshire, where increasing log sulphate levels seem to be associated with a linear increase in log TOC levels.

Figure 6.2.2 (c) shows the effect of increasing annual rainfall in the rivers located in Ayrshire – an increase in annual rainfall up to 1700mm appears to be associated with an increase in log TOC levels; however, a further increase appears to be associated with a decrease in log TOC levels. An increase in annual rainfall was associated with similar behaviour in the West Highland rivers.

Now, to consider the lochs located in the specified regions. Considering Table 6.2.2, highlights that 5 out of the 6 GAM models fitted to lochs in regions were improved with the inclusion of covariates, based on the adjusted R squared values – especially the lochs in Stirlingshire (increase from 60% to 83.1%) and the lochs in Lewis (increase from 42% to 80.4%). However, expression (6.2.1) was a more appropriate GAM model to be fitted to the lochs in the West Highlands. When fitting expression (6.2.1) to the lochs in the West Highlands, the only significant covariate was pH, which resulted in a decrease in the adjusted R-squared value from 83.5% to 43% (again, the missing pH values could be a plausible explanation for these results).

Summary of the Final GAM Models Fitted to Lochs in Regions													
Lochs	Region	Pr(> t )											Adjusted R. Sq
		Year	Month	Site	Year*Site	Month*Site	Temperature	pH	Log Alkalinity	Log Nitrate	Log Sulphate	Annual Rainfall	
	Dunbartonshire	<0.001	0.002	<0.001	<0.001	<0.001	0.04 (P)	-	-	-	0.008	-	29.5%
	West Highlands	0.03	<0.001	<0.001	<0.001	<0.001	-	-	-	-	-	-	83.5%
	Perthshire	<0.001	<0.001	<0.001	<0.001	<0.001	-	<0.001	<0.001 (P)	<0.001	<0.001		76.7%
	Stirlingshire	<0.001	<0.001	<0.001	<0.001	<0.001	0.002 (P)	<0.001	-	-	<0.001	<0.001	83.1%
	Sutherland	0.41 (P)	<0.001	<0.001	0.006	0.04	-	-	<0.001	-	<0.001		68.2%
	Lewis	0.09	0.02	<0.001	<0.001	<0.001	-	<0.001	<0.001	<0.001 (P)	-	<0.001 (p)	80.4%

**Table 6.2.2: Summary of the final GAM models fitted to lochs in the specified regions. Note: if a term was not included in the final GAM model, it is represented by “-”; all the terms in the final model are included as non-parametric terms, except from those marked with a “(P)” – these terms are parametric.**



**Figure 6.2.3: A selection of effect plots from the final GAM models fitted- pH in Stirlingshire (a); log sulphate in Stirlingshire (b); log alkalinity in Lewis (c); log nitrate in Perthshire (d).**

From studying Table 6.2.2, it is also clear that log sulphate is the most common covariate included in the final GAM models fitted to the lochs – log sulphate is included in 4 out of the 6 final GAM models fitted to the regions. Log alkalinity and pH are fairly common, as they are included in 3 out of the 6 final GAM models fitted. Annual rainfall is only significant in 2 out of the 6 regions. Even though, only 6 regions have been investigated, these results suggest that it is likely that log sulphate, log alkalinity and pH could also possibly explain the observed trends and patterns of log TOC in the regions of Scotland which were not considered. Figure 6.2.3 and 6.2.4 displays a selection of the effect plots from the final GAM models fitted to lochs in the different regions.

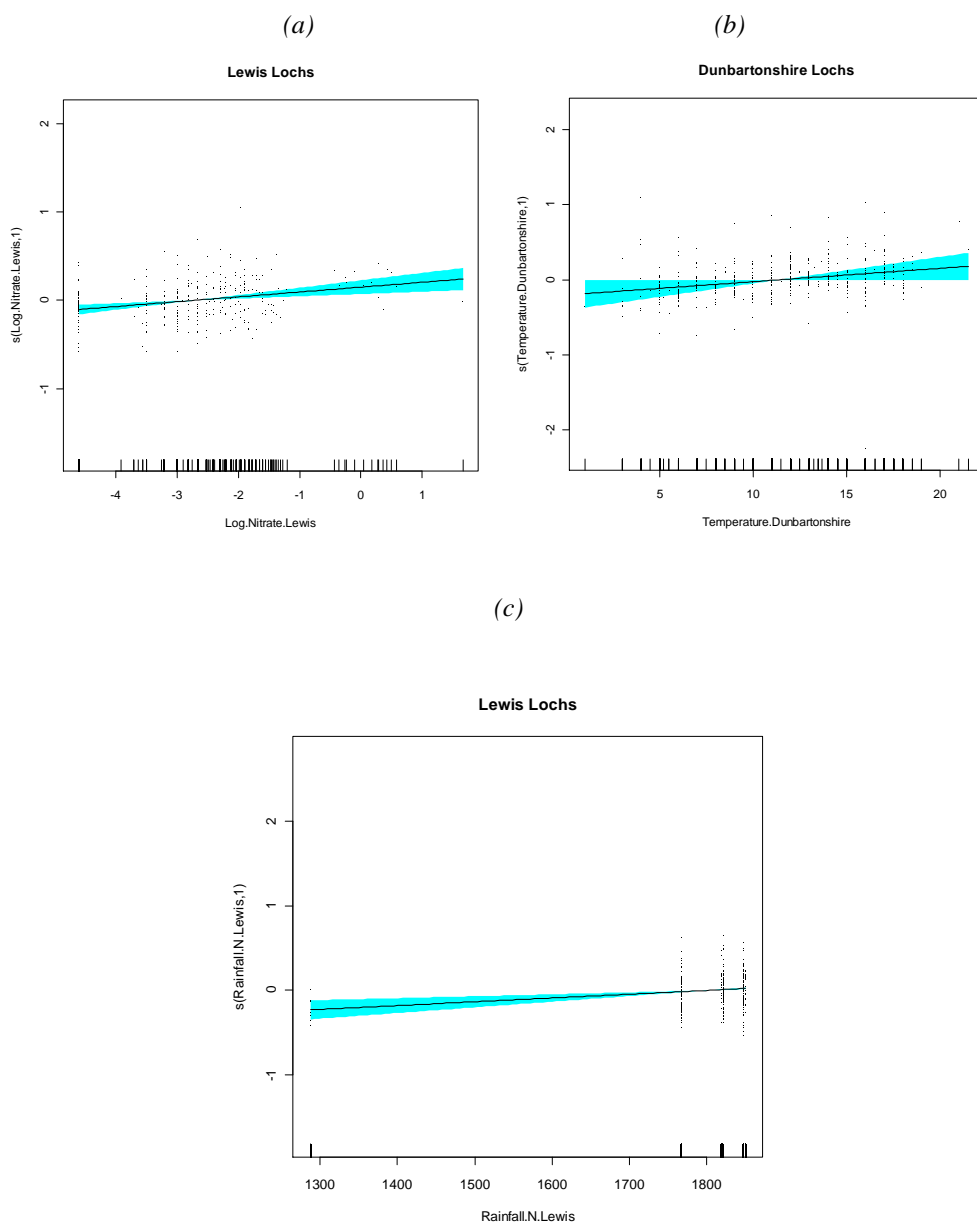
Figure 6.2.3 (a) displays the effect of increasing levels of pH in Stirlingshire – an effect which is similar in Lewis and Perthshire. An increase in pH levels between approximately 6 and 7 is associated with an increase in log TOC levels; however, if levels increase above 7, log TOC levels appear to level out.

Figure 6.2.3 (b) displays the effect of increasing levels of log sulphate in Stirlingshire. As levels of log sulphate increase to approximately 1 mg/l, levels of log TOC seem to decrease; but, as levels of log sulphate rise, levels of log TOC seem to increase rapidly. This behaviour is similar in Dunbartonshire and Sutherland. However, Perthshire displays a different pattern: as levels of log sulphate increase, levels of log TOC appear to smoothly decrease.

Figure 6.2.3 (c) shows the effect of increasing levels of log alkalinity in the lochs of Lewis. As levels of log alkalinity in the lochs increase, log TOC levels appear to increase rapidly – this behaviour is similar in Sutherland and Perthshire.

Figures 6.2.3 (d) and 6.2.4 (a) contrast the effects of increasing log nitrate levels in Perthshire and Lewis, respectively. An increase in log nitrate levels in Perthshire, is associated with a steady, smooth decrease in log TOC levels; however, in Lewis, it is associated with a linear increase in log TOC levels.

Figure 6.2.4 (b) shows the effect of increasing temperature levels in Dunbartonshire. Unlike the initial impressions gained in earlier chapters, Figure 6.2.4 (b) suggests that as the temperature increases, the levels of log TOC increases in a linear fashion.

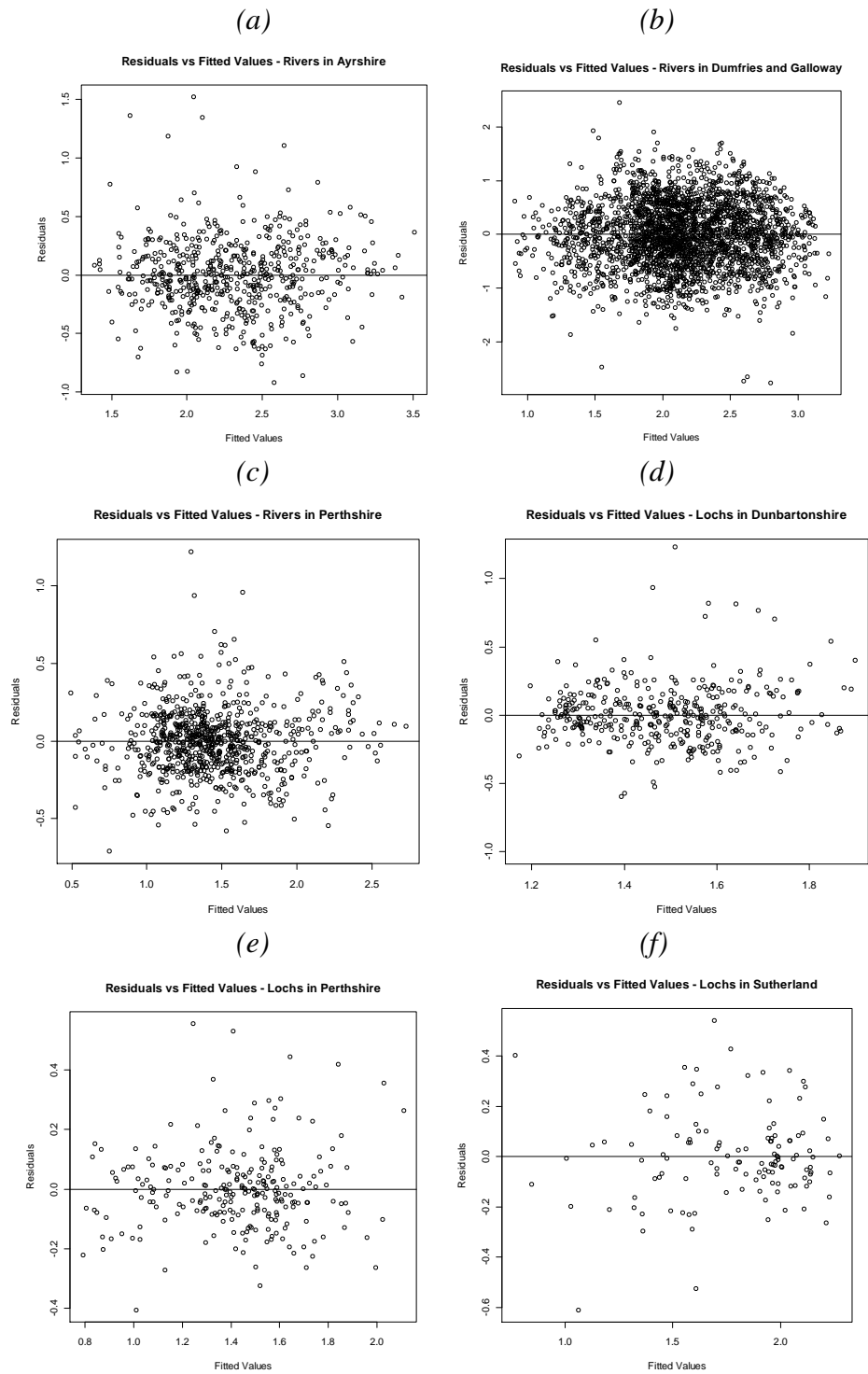


**Figure 6.2.4: A selection of effect plots from the final GAM models fitted- log nitrate in Lewis (a); temperature in Dunbartonshire (b); and annual rainfall in Lewis (c).**

Unlike the rivers, log TOC levels appear to continue to increase at temperature levels above 12 degrees Celsius. This behaviour is similar in the lochs in Stirlingshire.

Figure 6.2.4 (c) shows the effect of increasing annual rainfall in the lochs located in Lewis – an increase in annual rainfall appears to be associated with a linear increase in log TOC levels. An increase in annual rainfall was associated with similar behaviour in the Stirlingshire lochs.

In a usual manner, the validation of the final GAM models fitted to the regions (displayed in Tables 6.2.1 and 6.2.2) was considered by plotting the residuals against the fitted values. A selection of the residuals vs fitted values plots are displayed in Figure 6.2.5 – based on the plots, the residuals vs fitted values do not appear to show any signs of trends or patterns. Hence, the final GAM models fitted to each region seem to be appropriate.



**Figure 6.2.5:** A selection of residuals vs fitted values plots extracted from the final GAM models fitted in the regions Ayrshire (a), Dumfries and Galloway (b), and Perthshire (c) [with regards to rivers]; and the regions Dunbartonshire (d), Perthshire (e) and Sutherland (f) [with regards to lochs].



## 6.3 Conclusion

After exploring the coherency of sites in each region in the previous chapter, this chapter focused on fitting additive models to each region, which appropriately captured the log TOC levels over time and space. At first, the trend and seasonality of the regions were considered; but also, the interaction between the year and site, and the month and site. Tables 6.1.1 and 6.1.2 highlighted that the interaction terms were significant in all regions. This suggested that levels of log TOC differed between sites located in the same region. From the seasonal Mann-Kendall test performed in Chapter 5, it was found that sites located in the same region seem to have trends which are in the same direction; but, the direction of the trends are in a different direction in one or more seasons. However, the significant interaction terms in the GAM models suggest that even though the trends of the sites (in the same region) are in the same direction, the levels of log TOC vary from site to site across the years and throughout the months.

The significant interaction terms in the GAM models, required the results of the final DFA models in Chapter 5 to be investigated. All of the final DFA models fitted to the regions in Chapter 5 included only one underlying common trend. One common trend suggests that the levels of log TOC in the region are behaving very coherently, and that the levels do not significantly vary between sites. It has been established throughout the thesis that there is a strong seasonal pattern in the rivers and lochs, with regards to log TOC levels. Alonso et al., (2011) discussed seasonal dynamic factor analysis – unfortunately seasonal dynamic factor analysis could not be performed in the *Brodgar* software used for fitting the DFA models. Hence, the seasonal component of the data was removed as suggested by Zuur et al. (2004). The season seems to have a strong influence on the trend, and since this is not appropriately incorporated in the DFA models, it is possible that this will have affected the results of the final DFA models fitted to each of the regions.

After fitting GAM models which considered trend and seasonality, covariates were added to the models to try to improve the amount of variation explained in the data. It was found

that adding covariates to the trend and seasonality GAM models fitted to the rivers in the different regions, improved six out of the seven GAM models. It was only for the rivers in Argyll which the trend and seasonality GAM models was seen to be a more appropriate fit. Similarly, it was only the lochs in the West Highlands, where the trend and seasonality GAM model was a better fit to the data. None of the covariates fitted in the final GAM models were significant in all regions – for rivers or lochs. But, for the rivers, log alkalinity and log sulphate were significant in five out of the seven final GAM models fitted to the regions. With regards to the lochs, log sulphate was the most common covariate – it was included in four out of the six final GAM models fitted. Log alkalinity and pH were the second most common covariate fitted to the lochs as they were included in 3 out of the 6 final GAM models fitted to the regions.

# Chapter 7

## Discussions and Conclusions

### 7.1 Summary

The Scottish Environment Protection Agency (SEPA) is the regulatory agency responsible for monitoring water quality in Scottish waters and reporting back to the Scottish and UK governments and the European Community. The rising levels of organic carbon in Scottish rivers and lochs is of interest because it could indicate loss of soil carbon stocks and a source for carbon dioxide. Hence, the aim of this thesis was to perform a detailed investigation into the behaviour of total organic carbon and dissolved organic carbon. Furthermore, the aim was to explore physical and chemical factors which could possibly be driving such behaviour.

Having established that the thesis would focus on total organic carbon, due to the volume of missing dissolved organic carbon data, Chapter 2 explored trends, seasonality and relationships in rivers and lochs. The non-constant variability in total organic was clear in rivers and lochs; subsequently, a log transformation of the data suitably stabilized the

variability. The missing log TOC data at sites was also clear -data imputation was discussed, but the decision was made to work with the TOC data available as missing data would not present a problem for the standard regression techniques to be used in this thesis for analysis. Plotting the log TOC data for rivers and lochs, suggested that the trend was similar to the behaviour of dissolved organic carbon observed in the Northern Hemisphere, North America, central Europe and Scandinavia. Levels of log TOC appeared to increase throughout the 1990's, up until the early 2000's. It is not until the early 2000's that the increase in log TOC levels seems to weaken. Plotting the log TOC levels against the day of the year in which they were sampled, highlighted that there was a clear seasonal pattern in rivers and lochs: the log TOC levels seemed to be increasing from early spring until early autumn, which was followed by a steady decline through winter. However, the plots did suggest that the seasonal pattern appeared to be stronger in rivers.

Having explored the trend and seasonal patterns, the strength of the relationship between total organic carbon and dissolved carbon was of interest. Scatter plots and correlation tests suggested that there was a strong relationship between the two types of organic carbon. The relationships between log TOC and the covariates temperature, pH, alkalinity, nitrate, sulphate and river flow [data only available for 49 sites only] were also of interest. Similar to the ideas discussed in Chapter 1 by Freeman et al., (2001a) and Worrall et al., (2004), the plots suggested that temperature is associated with an increase in log TOC levels in Scottish rivers and lochs. Highest levels of log TOC are associated with a temperature of approximately 15 degrees Celsius. However, with regards to the effects of pH on log TOC, the effect seemed to be site specific, for both rivers and lochs. An increase in pH at one site is associated with an increase in log TOC; but, at other sites, it was the contrary. The site specificity was similar in rivers, with regards to log alkalinity's effect on log TOC; however, at loch sites, an increase in log alkalinity is associated with an increase in log TOC. The initial impression of log nitrate and log sulphate, suggested that unlike the other covariates, they do not seem to influence the levels of log TOC in either river or loch sites. The log TOC levels remain fairly flat, regardless of any increase or decrease in the log nitrate or sulphate concentration in the water. Based on visual exploration, for the sites with available flow data, it was found that an increase in the river flow is associated with an increase in log TOC levels at the sites.

The plotting of the different covariates raised two issues – values at the limit of detection and missing values. Log nitrate (in rivers and lochs) and log sulphate (in lochs) had values which were recorded at the limit of detection. To overcome this issue, a technique known as regression on order statistics (Helsel, 2005) was used, which seemed to effectively deal with the problem. Now, for the second issue – as temperature generally follows a seasonal pattern, the missing values could be predicted in a sensible manner by simple computation based on the monthly mean.

Chapter 2 provided an overview of the trends, seasonality and relationships in rivers and lochs; but, as 333 river sites and 187 loch sites were being considered, it was thought, that investigating individual sites in detail would be beneficial. Thus, Chapter 3 explored three river and three loch sites in detail. Sites with differing lengths of time series were chosen to represent the most common time periods in rivers and lochs. Plotting the trends, suggested, that for the sites with the longer time series (Callater Burn, Loch Kilbirnie, Loch Lomond), log TOC appears to increase up until the early 2000's, before "levelling off". The sites with data only between early 2000's and 2010 (River Tweed, River Dall Bridge, Loch Naver), did not show any strong trend – levels of log TOC remained fairly flat across the years. With regards to seasonality, log TOC seems to follow a seasonal pattern in all three river sites and Loch Kilbirnie. At these sites, it seems that levels of log TOC appear to increase from early spring up until early autumn – during late autumn and winter, the log TOC levels seem to decrease. There does not seem to be a strong seasonal pattern in either Loch Lomond or Loch Naver.

The relationship between log TOC and covariates were explored at each of the sites. At the river sites Tweed and Dall Bridge, an increase in the log Alkalinity levels was associated with a decrease in log TOC levels. An increase in temperature was associated with an increase in log TOC levels at each of the three sites. Log nitrate seemed to be associated with a decrease in log TOC levels at the River Tweed only. An increase in log flow seemed to be associated with an increase in log TOC at Callater Burn. On the other hand, the covariates did not appear to have a strong relationship with log TOC at any of the loch sites. If anything, an increase in temperature and log alkalinity seemed to be associated with an increase in log TOC – but, this was a very weak relationship. Based on the six sites investigated, the exploratory analysis suggested that the covariates were more likely to be associated with a

change in log TOC levels in rivers, than lochs. However, it was important to remember that only three river and three loch sites were being considered.

Having explored the trend, seasonality and relationship with covariates, Chapter 3 progressed on to considering different modelling techniques. Linear models and generalized additive models were explored – each model addressing trend, seasonality and the covariates. A linear model and generalized additive model was fitted to each site. It was found that the levels of log TOC at Callater Burn for any given month, on average, are increasing by 0.04 mg/l per year; and for any given month, on average, the level of log TOC is increasing by 0.02 mg/l at Loch Lomond (Creinch) and increasing by 0.06 mg/l at Loch Kilbirnie (Beith), per year. Moxley (2010), states that the rate of TOC increase, averaged across all Scottish sites with increasing concentrations, was 0.12 milligrams per litre per year (mg/l/y). Hence, the rate of increase does not seem to be as severe at these selected sites.

In Chapter 3, it was found that the length of time period did not seem to determine whether a linear or additive model was a more appropriate fit to a site. The river sites Callater Burn and River Tweed (with longer time series than the other site, Dall Bridge) were appropriately described by an additive model. This was expected, as the trends displayed by these sites, did not behave in a linear manner. However, Loch Lomond (Creinch) with the longest time series (out of the three lochs), was more appropriately described by a linear model. Based on these six sites, it seems that the most appropriate modeling technique is specific to each site.

The sites in Chapter 3 were not spatially or ecologically connected. Instead of continuing to explore sites on an individual basis, a logical next step was to consider sites which are connected in some manner. Chapter 4 considered sites which are located in, what has been described as, the River Dee network. The sites in the River Dee network were connected (or not connected) by the flow path of the river. In general, a river network consists of a main channel, and the streams and estuaries which flow in to the main channel. Therefore, a natural place to start was to consider five sites located on the main channel (River Dee itself). Initially, the sites were considered independently of one another. The exploratory analysis suggested that there was a common signal – the log TOC levels were increasing steadily until the early 2000's, which was followed by a weaker increase in the remaining years; there was a seasonal pattern evident in all sites; and the covariate 'log flow' seemed to influence log

TOC levels at all sites. Based on the exploratory analysis, the decision was made to not continue with further analysis of Banchory Bridge in Sections 4.2 and 4.3, due to the large amount of missing data. Similar to the previous chapter, two modelling approaches were used –linear and additive models were fitted to each site. Again, an approximate F-test was used to choose the “best” model to be fitted to each site. It was found that additive modelling was appropriate at three of the sites; and a linear model was more appropriate at Potarch Bridge.

Moving on from modelling each site separately, Chapter 4 attempted to find a global model to capture the behaviour of all four sites located on the main channel. To achieve this, a Generalized Additive Mixed Model (GAMM) was fitted to capture the common signals of the four sites. The final GAMM model revealed that there was a significant trend and seasonal pattern amongst the 4 sites; but also, the covariates log Alkalinity, log Sulphate and log Flow were associated with a change in log TOC levels at the four sites. The final GAMM model was more informative than the linear and additive models fitted to the sites individually. A global model was found to describe the behaviour of log TOC along the River Dee, which allowed the inclusion of a random site effect and a spatial correlation structure (exponential). In a sense, the inclusion of the spatial correlation structure highlighted that the distance between sites along the river, had an influence on the levels of log TOC.

Chapter 4 then progressed on to taking into consideration the sites located on the main channel, but also, the streams and estuaries flowing into the main channel. Defining the flow connectivity and the distance (Euclidean and river distance) between each of the sites, allowed the behaviour of log TOC across the network to be studied using a non-parametric smoothing technique developed by O'Donnell (2011). O'Donnell's smoothing technique effectively captured the structure of the network. At first, O'Donnell's technique was used to study the behaviour of log TOC over space –the log TOC values during the month of March in 2009 were chosen for analysis. O'Donnell's non-parametric smoothing technique was conducted using both river and Euclidean distance – it was found (regardless of which distance measurement was used) that as the river flows through the network, downstream towards site 1, the levels of log TOC seem to increase. Based on the visual inspection of

plots, and comparison of the root mean square error values, it was concluded, that river distance seems to be a more appropriate measurement between sites and would be used in the analysis to proceed. A natural progression from investigating the behaviour of log TOC over space was to monitor the trend of log TOC over time and space. To achieve this, four points in time were chosen – the log TOC values from March in the years: 1990, 1997, 2000 and 2009. The trend appeared to coincide with initial impressions previously formed in earlier sections – in the month of March (in the chosen years) the log TOC levels seemed to increase throughout the 1990's up until the early 2000's, and then “level off”. Levels of log TOC seem to increase between the years 1990 and 2000, particularly at the sites located where the river rises (near the Cairngorms).

The GAMM model appropriately captured the behaviour of sites situated on the same channel; however, in order to capture the common signals of the sites located across the network, a different approach was required. A GAM was fitted to initially capture the trend and seasonality of the log TOC levels across the network. The spatial location of the sites was included in the model as a covariate to capture the space element of the network; and the interactions between ‘year’ and ‘site’, and ‘month’ and ‘site’, were included, as it was thought that the levels of log TOC may be differ between sites. The covariates pH, temperature, log alkalinity, log nitrate, log sulphate and log flow were then added to the trend and seasonality GAM in an attempt to improve the model. It was found, that the “best” additive model to describe the log TOC levels of the thirteen sites in the River Dee network included: year, month, the interaction between the year and each site, the spatial location of the sites, log alkalinity, pH and log nitrate. The significant interaction between year and site, suggested that the levels of log TOC differ between the sites over the years - plotting the fitted values for each of the thirteen sites suggested that groups of sites (particularly sites close to each other) behaved coherently. Furthermore, this was supported by the significant spatial location term being included in the GAM – suggesting that the location of the site in the network influenced the levels of log TOC. However, the final GAM fitted to the River Dee network, was not a great fit to the data. Further research into the River Dee network could explore different environmental factors, such as the surrounding land use, which may be useful in explaining more of the variation in log TOC levels in the network.



Chapter 5 then addressed the main theme throughout the thesis – coherency. A literature review was conducted in Chapter 5 highlighting the variety of ways in which coherency has been measured in different papers. Having studied the literature, it was thought, that dynamic factor analysis and the seasonal Mann-Kendall test were appropriate techniques to be applied in this thesis. As the River Dee network was a key focus of Chapter 4, the coherency of the network sites was assessed. From the seasonal Mann-Kendall, it was found that the trend was in the same direction in each of the sites; but, the trend was only in the same direction during the season's winter, spring and autumn. The season seems to have a strong influence on the trend. Furthermore, the DFA highlighted that the best DFA model to describe the log TOC levels in the River Dee network included two common trends, the inclusion of a non-diagonal error covariance matrix and the explanatory variables annual mean temperature and annual rainfall. An overall interpretation of the coherency analysis of the network suggests that the log TOC at each of the network sites is behaving coherently; but, more specifically, there are actually two underlying common trends in the network. The annual mean temperature and annual rainfall in the east of Scotland appear to be driving the observed trends in the network.

Chapter 5 then considered rivers and lochs on a larger scale than the analysis carried out on the River Dee network - regions of Scotland were investigated. Based on exploratory analysis, it seemed plausible (for both rivers and lochs) that sites situated in the same region, have log TOC trends which could be described as being coherent. The trends displayed, supported previous subjective impressions of rivers: the log TOC levels seemed to increase up until the early 2000's, where the increase then either weakened or flattened out. However, exploring the trends of the lochs in different regions suggested that only in Dunbartonshire did log TOC levels behave similarly to the rivers regions. Previous exploratory analysis suggested that log TOC levels in lochs were fairly flat from 2005 onwards; but, analysis of each region highlighted that from 2005 onwards, the log TOC levels are fairly unsteady. The seasonality of log TOC within the regions was also considered (with regards to rivers and lochs) and was found to mirror the seasonal patterns seen previously.

Similar to the River Dee sites, a seasonal Mann-Kendall test was applied to a selection of the regions and DFA was performed to gain an understanding of the coherency of log TOC

levels in different sites located in the same region. The seasonal Mann-Kendall test was performed on the rivers in the regions West Highlands and Perthshire; and the lochs in the regions Lewis and Sutherland were considered. For each of these regions, it could be concluded that the trend of the sites was in the same direction; but, similar to the River Dee sites, the trend was not in the same direction in each of the seasons. Again, this re-iterates that the season could be a strong driver of trend in the regions. Chapter 5 then moved on to fitting dynamic factor analysis models to each of the regions – again, with varying number of common trends, the inclusion of an error covariance matrix (diagonal or non-diagonal) and the inclusion of explanatory variables common to all sites (annual mean temperature and annual rainfall). All of the final DFA models fitted included one common trend. This suggested that the log TOC levels of river and loch sites located in the same region, behave coherently. Also, nine out of the twelve regions studied, included either one or both of the explanatory variables in the final DFA models fitted. The environmental factors temperature and rainfall appear to influence the trends of log TOC in the majority of regions across Scotland.

After exploring the coherency of sites in Chapter 5, Chapter 6 focused on fitting additive models to each region, which appropriately captured the log TOC levels over time and space. At first, the trend and seasonality of the regions were considered; but also, the interaction between the year and site, and the month and site. It was found that the trend, seasonality, site and interaction between year and site, and month and site were all significant terms in the GAM's fitted to each region. The significant interaction terms in the GAM models suggest that even though the trends of the sites (in the same region) are in the same direction, the levels of log TOC vary from site to site across the years and throughout the months. However, the significant interaction terms in the GAM models did not seem to support the final DFA models fitted in Chapter 5. All of the final DFA models fitted to the regions in Chapter 5 included only one underlying common trend. One common trend suggests that the levels of log TOC in the region are behaving very coherently, and that the levels do not significantly vary between sites. However, it has been established throughout the thesis that there is a strong seasonal pattern in the rivers and lochs, with regards to log TOC levels – unfortunately seasonal dynamic factor analysis could not be performed in the *Brodgar* software used for fitting the DFA models. Hence, the seasonal component of the data was

removed as suggested by Zuur et al. (2004). The season seems to have a strong influence on the trend, and since this is not appropriately incorporated in the DFA models, it is possible that this will have affected the results of the final DFA models fitted to each of the regions.

Chapter 6 then focused on improving the trend and seasonality GAM models by including covariates. It was found that adding covariates to the trend and seasonality GAM models fitted to the rivers in the different regions, improved six out of the seven GAM models. It was only for the rivers in Argyll which the trend and seasonality GAM models was seen to be a more appropriate fit. Similarly, it was only the lochs in the West Highlands, where the trend and seasonality GAM model was a better fit to the data. None of the covariates fitted in the final GAM models were significant in all regions – for rivers or lochs. But, for the rivers, log alkalinity was significant in five out of the seven final GAM models fitted to the regions and pH and log sulphate were significant in four out of the seven. With regards to lochs, log sulphate was the most common covariate – it was included in four out of the six final GAM models fitted. Log alkalinity and pH were the second most common covariate fitted to the lochs as they were included in 3 out of the 6 final GAM models fitted to the regions.

## **7.2 Limitations of the Study and Future Work**

It is clear from the data provided by the Scottish Environment Protection Agency, for various reasons, total organic carbon samples were not obtained from the river and loch sites every month. Missing data did not present a problem for the regression techniques used throughout the thesis; however, a greater amount of data, may have displayed clearer trends and seasonal patterns of log TOC at particular sites. Furthermore, due to the location of sites (loch sites in particular), independent projects carried out at particular sites and lengths of time series, there was varying amount of log TOC data available for each of the river and loch sites. It is unrealistic to expect SEPA to have obtained the same number of total organic carbon samples for each of the 333 river and 187 loch sites over the past 30 years; but, the increasing awareness of the environment's wellbeing and the improvement in technology,

will hopefully lead to a greater sample size at each site, and allow for a fairer comparison of trends and patterns between sites, and provide more accurate results. Furthermore, the missing data restricted the analysis which could have been carried out on dissolved organic carbon. The relationship between log TOC and log DOC was explored at a selection of sites; however, given more available DOC data, it would have been interesting to compare the behaviour of log TOC across the River Dee network and regions of Scotland, to the behaviour of log DOC. Ideally, SEPA could increase the frequency of sampling at each, to gain a greater understanding of total organic carbon throughout each month, and over time. However, increased sampling could lead to analytical problems. Observations which are sampled days or weeks apart are more likely to be dependent and related to each other – this is an issue which would have to be addressed during analysis. Realistically, increased sampling frequency may not be cost effective, and may not improve the understanding of total organic carbon's behaviour significantly, to justify the cost. For the purpose of this thesis, missing data did not cause too many problems; but, future research into the behaviour of total organic carbon and dissolved organic carbon may want to consider exploring plausible techniques of imputing missing data in sensible and statistically sound manner.

Coherency was a theme at the heart of this thesis, one which was discussed in depth in Chapter 5. The dynamic factor analysis used in Chapter 5 was an effective measurement of coherency in the River Dee network, but also a selection of Scottish regions. Unfortunately, a seasonal component could not be incorporated into the dynamic factor analysis, using the software *Brodgar*. Due to the clear seasonal pattern evident in log TOC across Scottish rivers and lochs, it would be useful and appropriate, in future research, to be able to incorporate a seasonal component, as discussed by Alonso et al., (2011), and apply seasonal dynamic factor analysis to the River Dee network and regions across Scotland.

The main aim of Chapter 6 was appropriately capture behaviour of log TOC over time and space in a selection of Scottish regions. The  $\varepsilon_i$  in the GAM models fitted to each of the regions were assumed to be independent based on the spatial dependence analysis performed in Section 4.4.2. It was assumed that if the River Dee sites in the same network were deemed to be spatially independent, it seemed plausible that sites located in a larger geographical space, would also be spatially independent. Further research could explore the spatial dependence of sites located in the same region, taking into consideration that a region may

contain more than one river network; and also investigate alternative ways to capturing the ‘space’ element in regions – similar to Section 4.4.7, finding and including the spatial coordinates (longitude and latitude) of each site may have been a more appropriate way of capturing ‘space’.

The covariates temperature, pH, alkalinity, nitrate, sulphate and flow have been useful in explaining what is possibly driving the behaviour of log TOC in particular sites, the River Dee network and regions of Scotland; but, it is possible that other environmental factors could also be driving such behaviour. The environment is complex, and it seems more than likely, that several factors could be influencing the changes in total organic carbon. Further research should incorporate a wider spectrum of covariates to (hopefully!) improve the final models fitted to explain the behaviour of log TOC. For example, Worrall et al. (2007) and Clark et al. (2005), discussed the influence that changing water table depths had on DOC – incorporating the water table depth could improve the understanding of what is driving said trends and patterns of log TOC. Furthermore, Worrall et al. (2004), discussed the effect that changes in land management could have on DOC. Disturbances, such as afforestation, have been associated with short term increases in DOC in surrounding surface waters. Incorporating background information about the surrounding land management of rivers and lochs could be useful.

## **7.3 Conclusion**

The exploratory and formal analysis applied to the data, has indicated that in general, the log TOC levels has increased in Scottish rivers and lochs predominantly between the early 1990’s and early 2000’s. In the past five years, generally, an increase has not been observed in the rivers– the levels of log TOC have remained fairly constant. However, based on the regional analysis of lochs, from 2005 onwards, the log TOC levels appear to be fairly unsteady – showing signs of increasing and decreasing over the years. The analysis has also highlighted that log TOC appears to follow a seasonal pattern, although, it is more prevalent in rivers: log TOC levels seem to increase from early spring until early autumn, before

decreasing through winter. The dynamic factor analysis was effective in measuring the coherency of log TOC levels in regions – based on this method it seems plausible that river and loch sites located in the same region, are behaving coherently. Based on the final GAM models fitted to the regions, it seems plausible that the main drivers of change in log TOC levels are log alkalinity and log sulphate in rivers; and the main drivers of change in log TOC levels are log alkalinity, pH and log sulphate in lochs.

# Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716-723.

Alexander, L.V., Zhang X., Peterson T., Caesar J., Gleason B, Kelin T. A., Haylock M., Collins D., Trewin B., Rahimzadeh F., Tagipour A., Rupa Kumar K., Revadekar J., Griffiths A.M. Alonson., J.Rodriguez., C.G. Martos., M.J. Sanchez. ( 2011). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of geophysical research*. Vol. **111**.

Alonso, A. M., Rodriguez, J., Martos, C. G., Sanchez, M. J. (2011). Seasonal Dynamic Factor Analysis and Bootstrap Inference: Application to Electricity Market Forecasting. *Technometrics*, vol. **52**(2).

Baines, S. B., Webster, K. E., Kratz, T. K., Carpenter, S. R., Magnuson, J. J. (2000). Synchronous Behaviour of temperature, calcium and chlorophyll in lakes of northern Wisconsin. *Ecology*, **81**(3), pp. 815–825.

Barry, R. D. and Ver Hoef, J. M. (1996). Black box kriging: spatial prediction without specifying the variogram. *Journal of Agricultural, Biological, and Environmental Statistics* **1**(3).

Belle, G. v., Hughes, J. P. (1984). Non-parametric tests for trend in water quality. *Water Resources Research* **20**: 127-136.

Benson, B. J., Lenters, J. D., Magnuson, J. J., Stubbs, M. T., Kratz, K., Dillon, P. J., Hecky, R. E., and Lathrop, R. C. (2000). Regional coherence of climatic and lake thermal variables of four lake districts in the upper Great Lakes Region of North America. *Freshwater Biology* **43**:517–527.

Best, D. J. & Roberts, D. E. (1975), Algorithm AS 89: The Upper Tail Probabilities of Spearman's  $\rho$ . *Applied Statistics*, **24**, 377–379.

Blenckner, T., Adrian, R., Livingstone, D. M., Jennings, E., Weyhenmeyer, G.A., George, D.G., Jankowski, T., Jarvinen, M., Aonghusa, C.N., Noges, T., Strailess, D., Teubner, K. (2007). Large-scale climatic signatures in lakes across Europe: a meta-analysis. *Global Change Biology*, **13**, 1314-1326.

Bloch, I., Weyhenmeyer, G. A. (2010). Long-term changes in physical and chemical conditions of nutrient-poor lakes along a latitudinal gradient: is there a coherent phytoplankton community response? *Aquatic Science*.

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*. New York: John Wiley and Sons.

Bowman, A. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the kernel approach with S-Plus illustrations*, Oxford University Press.

Brillinger, D.R. (2001). *Time Series Data Analysis and Theory*. SIAM, Philadelphia.



Carey, S. K., Tetzlaff, D., Seibert, J., Soulsby, C., Buttle, J., Laudon, H., McDonnell J., McGuire, K., Caissie, D., Shanley, J., Kennedy, M., Devito, K., Pomeroy, J. W. (2010). Inter-comparison of hydro-climatic regimes across northern catchments: synchronicity, resistance and resilience. *Hydrol. Process.* **24**, 3591–3602

Chandler, R. E. and Scott, E. M. (2011). Statistical methods for Trend Detection and Analysis. John Wiley & Sons.

Chen, Z., Grasby, S. E. (2008). Impact of decadal and century-scale oscillations on hydroclimate trend analyses. *Journal of Hydrology* **365**, 122-133

Clarke, K. R., Green, R. H. (1988). Statistical design and analysis for a ‘biological effects’ study. *Mar Ecol Prog. Ser* **46**: 213–226.

Clark, J.M., Chapman, P.J., Adamson, J.K., Lane, S.J. (2005). Influence of drought-induced acidification on the mobility of dissolved organic carbon in peat soils. *Global Change Biology* **11**, 791–809.

Cressie, N. (1993), Statistics for Spatial Data, revised Edition, John Wiley and Sons.

Cressie, N., J. Frey, B. Harch, and M. Smith (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics* **11**(2), 127–150.

Curtis, C. E., Sun, F. T., Miller, L. M., D’Esposito, M. (2005). Coherence between fMRI time-series distinguishes two spatial working memory networks. *NeuroImage* **26**, 177 – 183.

Cygnus Research Interational. What is coherency function, phase function and gain function?

<http://www.cygres.com/OcnPageE/Glosry/Coh.html>

Dibiasi, A. & Bowman, A. (2001), 'On the use of the variogram in checking for independence in spatial data'. *Biometrics* **57**, 211-218.

Dixon & Turner, 1991. The global carbon cycle and climate change. *Environmental Pollution Volume 73, Issues 3-4*, Pages 245-262.

Driscoll, C.T., Driscoll, K.M., Roy, K.M., Mitchell, M.L. (2003). Chemical response of lakes in the Adirondack region of New York to declines in acidic deposition. *Bioscience* **61**(8).

Eimers, M. C., Watmough, S. A., Buttle, J. M. (2008). Long-term trends in dissolved organic carbon concentration: a cautionary note. *Biogeochemistry* **87**, 71-81.

Esterby, S. R. (1993). Trend Analysis Methods for Environmental Data. *Envirometrics* **4**, 459-481.

Environmental Science and Technology 37, 2036–2042. ECOSSE (2007). Estimating Carbon and Organic Soils Sequestration and Emissions. Scottish Executive.

Evans, C. D. & Montieth, D. T. (2001) Chemical trends at lakes and streams in the UK Acid Waters Monitoring Network, 1988-2000: evidence for recent recovery at a national scale. *Hydrol. Earth Sys. Sci.* **5**(3): 351-366

Evans, C. D., Freeman, C., Monteith, D. T., Reynolds, B. R., Fenner, N. (2002) Climate change – terrestrial export of organic carbon- reply. *Nature* **415**: 862-862

Fine, H. A., Dear, K. B. G., Loeffler, J. S., McBlack, P. L., Canellos, G. P. (1993). Meta-analysis of radiation therapy with and without adjuvant chemotherapy for malignant gliomas in adults. *Cancer* Volume **71**, Issue 8, pages 2585–2597.

Folster, J., Goransson, E., Johansson, K., A Wilander., 2005. Synchronous variation in water chemistry for 80 lakes in southern Sweden. *Environmental Monitoring and Assessment* **102**, 389–403.

Freeman C, Evans CD, Montieith DT, Reynolds B & Fenner N (2001a) Export of organic carbon from peat soils. *Nature* **412**: 785-785

George, D.G., Talling, J.F. and Rigg, E. (2000). Factors influencing the temporal coherence of five lakes in the English Lake District. *Freshwater*, **43**, 449-461.

Ghanbari, N. R., Bravol, H. R. (2011). Coherence Among Climate Signals, Precipitation, and Groundwater. *Groundwater* **49**(4), 476–490.

Ghanbari, R.N., Bravo, H.R., Magnuson, J.J., Hyzer, W.G., Benson, B.J. (2009). Coherence between lake ice cover, local climate, and teleconnections (Lake Mendota, Wisconsin). *Journal of Hydrology* **374**, no. 3–4: 282–293.

Gilbert, R. O. (1987). Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons.

Gremberghe, I. van., Wichelen, J. van., Gucht, K. van der., Vanormelingen, P., D'hondt, S., Boutte, C., Wilmotte, A., Vyverman, W. (2007). Covariation between zooplankton community composition and cyanobacterial community dynamics in Lake Blaarmeersen (Belgium). *FEMS Microbiol Ecol* **63**, 222–237.

Gurevitch, J., Morrison, J. A., Hedges, L. V. (2000). The interaction between competition and predation: a meta-analysis of field experiments. *American Naturalist* **155**, 435-453.

Hanson, R.T., Newhouse, M.W., Dettinger, M.D. (2004). A methodology to assess relations between climatic variability and variations in hydrologic time series in the southwestern United States. *Journal of Hydrology* **287**: 252–269.

Harvey, A.C. (1989). Forecasting, structural time series model and the Kalman filter. Cambridge University Press, Cambridge.

Hassan, M., Terrien, J., Karlsson, B., Marque, C. (2009). Application of wavelet coherence to the detection of uterine electrical activity synchronization in labor. *IRBM* doi:10.1016/j.irbm.2009.12.00

Hastie, T. and R. Tibshiranie (1986). Generalized additive models (with discussion). *Statistical Science* **1**, 297-318.

Hastie, T. and R. Tibshirani (1990). Generalised Additive Models. Chapman and Hall.

Hawkins, D. & Cressie, N. (1984). Robust kriging- a proposal. *Journal of the International Association of Mathematical Geology* **16**, 13-18 .

Hejzlar, J., Dubrovsky, M., Buchtele, J., Ruzicka, M. (2003). The apparent and potential effects of climate change on the inferred concentration of dissolved organic matter in a temperate stream (the Malse River, South Bohemia). *Science of the Total Environment* **310**, 143–152.

Helsel, D., Hirsch, R. (1992). Statistical methods in water resources. Studies in environmental science. Elsevier, Amsterdam. *Environmental Science* **49**.

Helsel, D. R. (2005). Nondetects And Data Analysis: John Wiley and Sons, New York.

Helsel, Lee. (2005). Statistical analysis of environmental data containing multiple detection limits: S-language software for regression on order statistics, *Computers in Geoscience* vol. **31**, pp. 1241-1248.

Helsel, D. R. (2005). Nondetects And Data Analysis: Statistics for censored environmental data. John Wiley and Sons, New York. 250p.

Hirsch, R. M., Slack, J. R., Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research* **18**: 107-121.

Hoef, Ver., Peterson, E., Theobald, D. (2006). Spatial statistical models that uses flow and stream distance. *Environ Ecol Stat* **13**:449-464 .

Hollander, M. & Wolfe, D. A. (1973), *Nonparametric Statistical Methods*. New York: John Wiley & Sons. Pages 185–194 (Kendall and Spearman tests).

Hudgins, L., Friehe, A., Mayer, M. (1993). Wavelet transforms and atmospheric turbulence. *Phys. Rev. Lett.* **71**,3279–82.

Jenkins, G.M., and Watts, D.G. (1968). Spectral Analysis and Its Applications. San Francisco: Holden-Day.

Kendall, M.G. (1975). Rank Correlation Methods, 4<sup>th</sup> ed. Charles Griffin, London.

- Kent, A.D., Yannarell, A. C., Rusak, J. A., Triplett, E. W., McMahon, K. D. (2007). Synchrony in aquatic microbial community dynamics. *The ISME Journal* **1**, 38–47
- Kiktev, D., Sexton, D. M., Alexander, L., Folland, C. (2003). Comparison of modeled and observed trends in indices of daily climate extremes. *Journal of Climate* **16**: 3560–3571.
- Krug, E. C. and Frink, C. R. (1983). Acid rain on acid soil: a new perspective. *Science* **211**: 520-525
- Lachaux, J., Lutz, A., Rudrauf, D., Cosmelli, D., Le Van Quyen, M., Martinerie, J., Varela, F. (2002). Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiol. Clin.* **32**, 157–74.
- Lange, H., Bernhardt, K. (2004). Long term components and regional synchronization of river run offs. *Hydrology: Science and Practice for the 21st century*. Volume I.
- Legendre, P., Legendre, L. (1998). Numerical Ecology. Elsevier Science: BV, Amsterdam.
- Lopes, H.F., Salazar, E., Gamerman, D. (2008). Spatial Dynamic Factor Analysis. *Bayesian Analysis* **3**, Number 4, pp. 759-792.
- Lunneborg C. E. (2000). Data Analysis by Resampling – Concepts and Applications. Duxbury Press: Pacific Grove, CA.
- Maraun, D., Kurths, J. (2004). Cross wavelet analysis: significance testing and pitfalls. *Nonlin. Proces. Geophys.* **11**, 505–14.

Magnuson, J. J., Kratz, T. K., Benson, B. J., Webster, K. E. (2006b). Coherent dynamics among lakes. Pages 89–106 in J. J. Magnuson, T. K. Kratz, and B. J. Benson, editors. Longterm dynamics of lakes in the landscape. Oxford University Press, Oxford, UK.

Magnuson, J.J., Benson, B.J., Kratz, T.K. (1990). Temporal coherence in the limnology of a suite of lakes in Wisconsin, U.S.A. *Freshwater Biology* **23**, 145-159.

Mann, H.B. (1945). Non-parametric tests against trend. *Econometrica* **13**:245-259.

Marshall, D., Johnell, O., Wedel, H. (1996). Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *BMJ* 18;312(7041):1254-9.

Monteith, D., Stoddard, J. L., Evans, J. L. C. D, De Wit, H. A., Forsius, M., Hogasen, T., Wilander, A., Skjelkvale, B. L., Jeffries, S. D., Vuorenmaa, J., Keller, B., Kopacek, J., and Vesely, J. (2007). Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry. *Nature* **450**: 537-540

Monk, W. A., Wood, P. J., Hannah, D. M., Wilson, D. A. (2007). Selection of river flow indices for the assessment of hydroecological change. *River Research and Applications* **23**: 113–122.

Moxley. J. (2011). Trends in organic carbon in Scottish rivers and lochs. Scottish Environment Protection Agency.

Muller, K., Lohmann, G., Bosch, V., von Cramon, D.Y. (2001). On multivariate spectral analysis of fMRI time series. *NeuroImage* **14**, 347– 356.

Munoz-Carpena, R., Ritter, A., Li, Y.C. (2005). Dynamic factor analysis of ground water quality trends in an agricultural area adjacent to Everglades National Park. *Journal of Contaminant Hydrology*. Volume: **80**, Issue: 1-2, Pages: 49-70

Nash, J.E., Sutcliffe, J.V. (1996). River flow forecasting through conceptual models. Part 1- A discussing of Principles. *J. Hydrol.* **10**, 282-290.

Nye, J. A., Bundy, A., Shackell, N., Friedland, K. D., Link, J. S. (2008). Coherent trends in contiguous survey time-series of major ecological and commercial fish species in the Gulf of Maine ecosystems. *ICES Journal of Marine Science*, **67**: 26–40

O'Donnell, D. (2011). Spatial prediction and spatial temporal modelling on river networks. PhD thesis. University of Glasgow.

Pace, M. L., Cole, J. J. (2002). Synchronous variation of dissolved organic carbon and color in lakes. *Limnol. Oceanogr.* **47**(2), 333–342.

Patoine, A., Leavitt, P. R. (2006). Century-long synchrony of fossil algae in a chain of Canadian prairie lakes. *Ecology*, **87**(7), pp. 1710–1721

Potamias, G., Moustakis, V. S. (2001). Mining coherence in time series data in: *Proceedings of the Ninth International Conference on Human-Computer Interaction* pp. 928-932

Pfister, L., Humbert, J., Hoffman, L. (2000). Recent trends in rainfall-runoff characteristics in the Alzette river basin, Luxembourg. *Climatic Change* **45**: 323–337.

Piegorsch, W. W., Bailer, A. J. (2005). Analyzing environmental data. John Wiley & Sons.



Polansky, L., Wittemyer, G., Cross, P. C., Tambling, C. J. (2010). From moonlight to movement and synchronized randomness: Fourier and wavelet analyses of animal location time series data. *Ecology*, **91**(5), 2010, pp. 1506–1518

Pryor, S. C., However, J. A., Kunkel, K. E. (2009). How spatially coherent and statistically robust are temporal changes in extreme precipitation in the contiguous USA? *Int. J. Climatol.* **29**: 31–45 (2009)

Rincon, F. A. (2009). Environmental trends over both time and space. MSc thesis. University of Glasgow.

River Dee Map: <http://www.theriverdee.org/explore-the-catchment.asp>

Rosenberg, M. S., Adams, D. C., Gurevitch, J. (2000). MetaWin Version 2.0 Statistical Software for Meta-Analysis. Sinauer, Sunderland.

Saitou, N., Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** (4), 406–425

Sanderson, J., Fryzlewicz, P., Jones, M. W. (2010). Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, **97**, 2, pp. 435–446.

Sen, P.K. (1968b). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* **63**: 1379-1389.

SEPA Chemistry<sup>1</sup>: Determination of TOC and DOC in fresh and waste waters using sodium persulphate by Aurora 1030 W. ES-INR-P-004

SEPA Chemistry<sup>2</sup>: National Inorganics Work Procedure. Determination of TOC and DOC in fresh waters and effluents using high temperature catalysed combustion analysis. ES-INR-P-708.

SEPA chemistry <sup>3</sup>: Determination of pH, conductivity and alkalinity by radiometer Titrab Tim 900. ES-INR-P-900-pH-alk

SEPA chemistry <sup>4</sup>: Determination of pH, electrical conductivity and alkalinity by Metrohm Autotitrator. ES-INR-P-800-pH-alk

SEPA (2009-2015). River basin management plan for the Scotland river basin districts. Technical report.

Skjelkvale, B.L., Mannio, J., Wilander, A., Andersen, T. (2001). Recovery from acidification of lakes in Finland, Norway and Sweden 1990–1999. *Hydrology and Earth System Science* **5**, 327–338.

Sniffer (2006). Handbook of Climate Trends Across Scotland. UKCP09 scenarios.

Stoddard, J.L., Karl, J.S., Devinev, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoc, P.S., Webb, J.R., Webster, K.E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Report EPA 620/R-03/001.

Strickland, C. M., Simpson, D. P., Turner, I. W., Denham, R. Mengersen, K. L. (2009). Fast Bayesian analysis of spatial dynamic factor models for large space time data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. Volume **60**, Issue 1, pages 109–124.

Tetzlaff, D., Soulsby, C., Birkel, C. (2010). Hydrological connectivity and microbiological fluxes between landscapes and riverscapes: the importance of seasonality. *Hydrological Processes* **24**: 1231–1235.

Torrence, C. and Webster, P. J. (1998). A practical guise to wavelet analysis. *Bulletin of the American Meteorological Society*, **79**(1)67-78.

Tranvik, L. J. & Jansson, M. (2002) Terrestrial export of organic carbon. *Nature* **415**: 861-862.

Urban Waste Water Treatment Directive (1991). Directive 91/71/EEC Council Directive 21 May 1991 concerning urban waste water treatment.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition Springer.

Ver Hoef, J. M., E. Peterson, and D. Theobald (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* **13**(4), 449–464.

Water Framework Directive (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000, establishing a framework for community action in the field of water policy.

Weyhenmeyer, G. A. (2008). Water chemical changes along a latitudinal gradient in relation to climate and atmospheric deposition. *Climatic Change*, **88**:199–208.

Weyhenmeyer, G. A. and Karlsson, J. (2009). Nonlinear response of dissolved organic carbon concentrations in boreal lakes to increasing temperatures. *Limnol. Oceanogr.* **54** (6, part 2), 2513–2519

Webster, R. and M. A. Oliver (2001). *Geostatistics for Environmental Scientists*. Wiley.

The EU Water Framework Directive - integrated river basin management for Europe, [http://ec.europa.eu/environment/water/water-framework/index\\_en.html](http://ec.europa.eu/environment/water/water-framework/index_en.html)

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester, UK.

Wood, S. N. (2006). *Generalized Additive Models. An introduction with R*. Chapman and Hall/CRC.

Worrall, F., Burt, T.P., Shedden, R. (2003). Long terms records of riverine carbon flux. *Biogeochemistry* **64**, 165–178.

Worrall, F., Harriman, R., Evans, C. D., Watts, C. D., Adamson, J., Neal, D., Tipping, G., Burt, T., Grieve, I., Monteith, I., Nadeen, P. S., Nisbet, T., Reynolds, B., Steves, P. (2004) Trends in dissolved organic carbon in UK rivers and lakes. *Biogeochemistry* **70**: 369-402.

Worrall, F., Burt, T. P. (2007). Trends in DOC concentration in Great Britain. *Journal of Hydrology*, Volume **346**, Issues 3-4, Pages 81-92.

Zuur, A.F., Tuck, I.D., Bailey, N. (2003a). Dynamic factor analysis to estimate common trends in fisheries time series. *Can. J. Fish. Aquat. Sci* **60**, 542-552

Zuur, A.F., Fryer, R.J., Jolliffe, I.T., Dekker, R., Beukema, J.J. (2003b). Estimating common trends in multivariate time series using dynamic factor analysis. *Envirometrics* **14** (7), 665-685.

Zuur, A.F., Pierce, G.J. (2004). Common trends in northeast Atlantic squid time series. *J. Sea Res.* **52**, 57-72.

Zuur, A.F., Leno, E. N., Smith, G. M. (2007). *Analysing Ecological Data*. Springer.

Zuur, A. (2011). Highland Statistics Ltd. [www.brodgar.com](http://www.brodgar.com) (Version 2.7.2).